

Lossless Data Compression at Finite Blocklengths

Ioannis Kontoyiannis

Department of Informatics

Athens University of Economics and Business

Athens 10434, Greece

yiannis@aueb.gr

Sergio Verdú

Department of Electrical Engineering

Princeton University

Princeton, New Jersey 08544, USA

verdu@princeton.edu

Abstract

This paper provides an extensive study of the behavior of the best achievable rate (and other related fundamental limits) in variable-length lossless compression. In the non-asymptotic regime, the fundamental limits of fixed-to-variable lossless compression with and without prefix constraints are shown to be tightly coupled. Several precise, quantitative bounds are derived, connecting the distribution of the optimal codelengths to the source information spectrum, and an exact analysis of the best achievable rate for arbitrary sources is given.

Fine asymptotic results are proved for arbitrary (not necessarily prefix) compressors on general mixing sources. Non-asymptotic, explicit Gaussian approximation bounds are established for the best achievable rate on Markov sources. The source dispersion and the source varentropy rate are defined and characterized. Together with the entropy rate, the varentropy rate serves to tightly approximate the fundamental non-asymptotic limits of fixed-to-variable compression for all but very small blocklengths.

Keywords — Lossless data compression, fixed-to-variable source coding, fixed-to-fixed source coding, entropy, finite-blocklength fundamental limits, central limit theorem, Markov sources, varentropy, minimal coding variance, source dispersion.

I. FUNDAMENTAL LIMITS

A. The optimum fixed-to-variable code

A fixed-to-variable compressor for a finite alphabet \mathcal{A} is an injective function,

$$f_n: \mathcal{A}^n \rightarrow \{0, 1\}^* = \{\emptyset, 0, 1, 00, 01, 10, 11, 000, 001, \dots\}. \quad (1)$$

The length of a string $a \in \{0, 1\}^*$ is denoted by $\ell(a)$. Therefore, a block (or file) of n symbols $a^n = (a_1, a_2, \dots, a_n) \in \mathcal{A}^n$ is losslessly compressed by f_n into a binary string whose length is $\ell(f_n(a^n))$ bits.

When the file $X^n = (X_1, X_2, \dots, X_n)$ to be compressed is generated by a probability law P_{X^n} , a basic information-theoretic object of study is the distribution of the rate of the optimal compressor, seen as a function of the blocklength n and the distribution P_{X^n} . The best achievable compression performance at finite blocklengths is characterized by fundamental limits, including:

- 1) $R^*(n, \epsilon)$: The lowest rate R such that the compression rate of the best code exceeds R with probability not greater than ϵ :

$$\min_{f_n} \mathbb{P}[\ell(f_n(X^n)) > nR] \leq \epsilon. \quad (2)$$

- 2) $\epsilon^*(n, k)$: The smallest possible excess-rate probability, namely, the probability that the compressed length is greater than or equal to k :

$$\epsilon^*(n, k) = \min_{f_n} \mathbb{P}[\ell(f_n(X^n)) \geq k]. \quad (3)$$

- 3) $n^*(R, \epsilon)$: The smallest blocklength at which compression at rate R is possible with probability at least $1 - \epsilon$; in other words, the minimum n required for (2) to hold.
- 4) $\bar{R}(n)$: The minimal average compression rate:

$$\bar{R}(n) = \frac{1}{n} \min_{f_n} \mathbb{E}[\ell(f_n(X^n))] \quad (4)$$

$$= \frac{1}{n} \sum_{k=1}^{\infty} \epsilon^*(n, k). \quad (5)$$

Naturally, the fundamental limits in 1), 2) and 3) are equivalent in the sense that knowledge of one of them (as a function of its parameters) determines the other two. For example,

$$R^*(n, \epsilon) = \frac{k}{n} \text{ if and only if } \epsilon^*(n, k) \leq \epsilon < \epsilon^*(n, k-1). \quad (6)$$

As for 4), we observe that, together with (5) and the fact that $\epsilon^*(n, 0) = 1$, (6) results in:

$$\bar{R}(n) = \int_0^1 R^*(n, x) dx - \frac{1}{n}. \quad (7)$$

The minima in the fundamental limits (2), (3), (4) are achieved by an optimal compressor f_n^* that assigns the elements of \mathcal{A}^n ordered in decreasing probabilities to the elements in $\{0, 1\}^*$ ordered lexicographically as in (1). In particular,

$$\bar{R}(n) = \frac{1}{n} \mathbb{E}[\ell(f_n^*(X^n))], \quad (8)$$

and,

$$R^*(n, \epsilon) \text{ is the smallest } R \text{ s.t. } \mathbb{P}[\ell(f_n^*(X^n)) > nR] \leq \epsilon, \quad (9)$$

where the optimal compressor f_n^* is described precisely as:

Property 1: For every $k = 1, \dots, \lfloor \log_2(1 + |\mathcal{A}|^n) \rfloor$, any optimal code f_n^* assigns strings of length $0, 1, 2, \dots, k-1$ to each of the

$$1 + 2 + 4 + \dots + 2^{k-1} = 2^k - 1, \quad (10)$$

most likely elements of \mathcal{A}^n . If $\log_2(1 + |\mathcal{A}|^n)$ is not an integer, then f_n^* assigns strings of length $\lfloor \log_2(1 + |\mathcal{A}|^n) \rfloor$ to the least likely $|\mathcal{A}|^n + 1 - 2^{\lfloor \log_2(1 + |\mathcal{A}|^n) \rfloor}$ elements in \mathcal{A}^n .

Note that Property 1 is a necessary and sufficient condition for optimality, which does not determine f_n^* uniquely: not only does it not specify how to break ties among probabilities but any swap between two codewords of the same length preserves optimality. As in the following example, it is convenient, however, to think of f_n^* as the unique compressor constructed by breaking ties lexicographically and by assigning the elements of $\{0, 1\}^*$ in the lexicographic order of (1).

Example 1: Suppose $n = 4$, $\mathcal{A} = \{\circ, \bullet\}$, and the source is memoryless with $\mathbb{P}[X = \bullet] > \mathbb{P}[X = \circ]$. Then the following compressor is optimal:

$$\begin{aligned} f_4^*(\bullet \bullet \bullet \bullet) &= \emptyset \\ f_4^*(\bullet \bullet \bullet \circ) &= 0 \\ f_4^*(\bullet \bullet \circ \bullet) &= 1 \\ f_4^*(\bullet \circ \bullet \bullet) &= 00 \\ f_4^*(\circ \bullet \bullet \bullet) &= 01 \\ f_4^*(\circ \circ \bullet \bullet) &= 10 \\ f_4^*(\circ \bullet \bullet \circ) &= 11 \\ f_4^*(\circ \bullet \circ \bullet) &= 000 \\ f_4^*(\bullet \bullet \circ \circ) &= 001 \\ f_4^*(\bullet \circ \circ \bullet) &= 010 \\ f_4^*(\bullet \circ \bullet \circ) &= 011 \\ f_4^*(\circ \circ \circ \bullet) &= 100 \\ f_4^*(\circ \circ \bullet \circ) &= 101 \\ f_4^*(\circ \bullet \circ \circ) &= 110 \\ f_4^*(\bullet \circ \circ \circ) &= 111 \\ f_4^*(\circ \circ \circ \circ) &= 0000. \end{aligned}$$

We emphasize that the optimum code f_n^* is independent of the design target, in that, e.g., it is the same regardless of whether we want to minimize average length or the probability that the encoded length exceeds 1 KB or 1 MB. In fact, the code f_n^* possesses the following

strong stochastic (competitive) optimality property over any other code f_n that can be losslessly decoded:

$$\mathbb{P}[\ell(f_n(X^n)) \geq k] \geq \mathbb{P}[\ell(f_n^*(X^n)) \geq k], \quad \text{for all } k \geq 0. \quad (11)$$

Note that, although f_n^* is not a prefix code, the decompressor is able to recover the source file a^n exactly from $f_n^*(a^n)$ and its knowledge of n and P_{X^n} . Since the whole source file is compressed, it is not necessary to impose a prefix condition in order for the decompressor to know where the compressed file starts and ends. Removing the prefix-free constraint at the block level, which is extraneous in most applications, results in higher compression efficiency.

B. Optimum fixed-to-variable prefix codes

The fixed-to-variable prefix code that minimizes the average length is the Huffman code, achieving the average compression rate $\bar{R}_p(n)$ (which is strictly larger than $\bar{R}(n)$), defined as in (4) but restricting the minimization to prefix codes. Alternatively, as in (2), we can investigate the optimum rate of the prefix code that minimizes the probability that the length exceeds a given threshold. If the minimization in (2) is carried out with respect to codes that satisfy the prefix condition then the corresponding fundamental limit is denoted by $R_p(n, \epsilon)$, and analogously $\epsilon_p(n, k)$ for (3). Note that the optimum prefix code achieving the minimum in (3) will, in general, depend on k . The following result shows that the corresponding fundamental limits, with and without the prefix condition, are tightly coupled:

Theorem 1: Suppose all elements in \mathcal{A} have positive probability. For all $n = 1, 2, \dots$

1) For each $k = 1, 2, \dots$:

$$\epsilon_p(n, k+1) = \begin{cases} \epsilon^*(n, k) & k < n \log_2 |\mathcal{A}| \\ 0 & k \geq n \log_2 |\mathcal{A}|. \end{cases} \quad (12)$$

2) If $|\mathcal{A}|$ is not a power of 2, then for $0 \leq \epsilon < 1$:

$$R_p(n, \epsilon) = R^*(n, \epsilon) + \frac{1}{n}. \quad (13)$$

If $|\mathcal{A}|$ is a power of 2, then (13) holds for $\epsilon \geq \min_{a^n \in \mathcal{A}^n} P_{X^n}(a^n)$, while we have,

$$R_p(n, \epsilon) = R^*(n, \epsilon) = \log_2 |\mathcal{A}| + \frac{1}{n}, \quad (14)$$

for $0 \leq \epsilon < \min_{a^n \in \mathcal{A}^n} P_{X^n}(a^n)$.

Proof: 1): fix k and n satisfying $2^k < |\mathcal{A}|^n$. Since there is no benefit in assigning shorter lengths, any Kraft-inequality-compliant code f_n^p that minimizes $\mathbb{P}[\ell(f_n(X^n)) > k]$ assigns length k to each of the $2^k - 1$ largest masses of P_{X^n} . Assigning all the other elements in \mathcal{A}^n lengths equal to

$$\ell_{\max} = \lceil k + \log_2(|\mathcal{A}|^n - 2^k + 1) \rceil, \quad (15)$$

guarantees that the Kraft sum is satisfied. On the other hand, according to Property 1, the optimum code f_n^* without prefix constraints encodes each of the $2^k - 1$ largest masses of P_{X^n} with lengths ranging from 0 to $k - 1$. Therefore,

$$\mathbb{P}[\ell(f_n^p(X^n)) \geq k + 1] = \mathbb{P}[\ell(f_n^*(X^n)) \geq k]. \quad (16)$$

Alternatively, if $2^k \geq |\mathcal{A}|^n$, then a zero-error n -to- k code exists, and therefore $\epsilon_p(n, k+1) = 0$.

2): According to f_n^* the length of the longest codeword is $\lceil n \log_2 |\mathcal{A}| \rceil$. Therefore,

$$\epsilon^*(n, \lceil n \log_2 |\mathcal{A}| \rceil + 1) = 0 \quad (17)$$

and

$$\epsilon^*(n, \lceil n \log_2 |\mathcal{A}| \rceil) = \begin{cases} 0 & |\mathcal{A}| \text{ is not a power of 2} \\ \min_{a^n \in \mathcal{A}^n} P_{X^n}(a^n) & |\mathcal{A}| \text{ is a power of 2} \end{cases} \quad (18)$$

On the other hand, 1) implies

$$\epsilon_p(n, \lceil n \log_2 |\mathcal{A}| \rceil + 1) = 0 \quad (19)$$

$$\epsilon_p(n, \lceil n \log_2 |\mathcal{A}| \rceil) = \epsilon^*(n, \lceil n \log_2 |\mathcal{A}| \rceil - 1) \quad (20)$$

$$\geq \min_{a^n \in \mathcal{A}^n} P_{X^n}(a^n) \quad (21)$$

Furthermore, $R_p(n, \cdot)$ can be obtained from $\epsilon_p(n, \cdot)$ through the counterpart to (6):

$$R_p(n, \epsilon) = \frac{i}{n} \text{ if and only if } \epsilon_p(n, i) \leq \epsilon < \epsilon_p(n, i+1). \quad (22)$$

Together with (6) and (12), (22) implies that (13) holds if $\epsilon \geq \min_{a^n \in \mathcal{A}^n} P_{X^n}(a^n)$. Otherwise, (18)-(21) result in (14) when $|\mathcal{A}|$ is a power of 2. If $|\mathcal{A}|$ is not a power of 2 and $0 \leq \epsilon < \min_{a^n \in \mathcal{A}^n} P_{X^n}(a^n)$, then

$$R^*(n, \epsilon) = \frac{\lceil n \log_2 |\mathcal{A}| \rceil}{n} \quad (23)$$

$$R_p(n, \epsilon) = \frac{\lceil n \log_2 |\mathcal{A}| \rceil + 1}{n} \quad (24)$$

■

C. The optimum fixed-to-fixed almost-lossless code

As pointed out in [30], [31], the quantity $\epsilon^*(n, k)$ is, in fact, intimately related to the problem of almost-lossless fixed-to-fixed data compression. Assume the nontrivial compression regime in which $2^k < |\mathcal{A}|^n$. The optimal n -to- k fixed-to-fixed compressor assigns a unique string of length k to each of the $2^k - 1$ most likely elements of \mathcal{A}^n , and assigns all the others to the remaining binary string of length k , which signals an *encoder failure*. Thus, the source strings that are decodable error-free by the optimal n -to- k scheme are precisely those that are encoded with lengths ranging from 0 to $k-1$ by the optimum variable-length code (Property 1). Therefore, $\epsilon^*(n, k)$, defined in (3) as a fundamental limit of (strictly) lossless variable-length codes is, in fact, equal to the minimum error probability of an n -to- k code. Accordingly, the results obtained in this paper apply to the standard paradigm of almost-lossless fixed-to-fixed compression as well as to the setup of lossless fixed-to-variable compression without prefix-free constraints at the block level.

The case $2^k \geq |\mathcal{A}|^n$ is rather trivial: the minimal probability of encoding failure for an n -to- k code is 0, which again coincides with $\epsilon^*(n, k)$, unless $|\mathcal{A}|^n = 2^k$, in which case, as we saw in (18),

$$\epsilon^*(n, k) = \min_{a^n \in \mathcal{A}^n} P_{X^n}(a^n). \quad (25)$$

D. Existing asymptotic results

Based on the correspondence between almost-lossless fixed-to-fixed codes and prefix-free lossless fixed-to-variable codes, previous results on the asymptotics of fixed-to-fixed compression can be brought to bear. In particular the Shannon-MacMillan theorem [24], [17] implies that for a stationary ergodic finite-alphabet source with entropy rate H , and for all $0 < \epsilon < 1$,

$$\lim_{n \rightarrow \infty} R^*(n, \epsilon) = H. \quad (26)$$

It follows immediately from Theorem 1 that the prefix-free condition incurs no loss as far as the limit in (26) is concerned:

$$\lim_{n \rightarrow \infty} R_p(n, \epsilon) = H, \quad (27)$$

Suppose X^n is generated by a memoryless source with distribution

$$P_{X^n} = P_X \times P_X \times \cdots \times P_X, \quad (28)$$

and define the *information random variable*,¹

$$\iota_X(X) = \log_2 \frac{1}{P_X(X)}. \quad (29)$$

For the expected length, Szpankowski and Verdú [27] show that the behavior of (4) for non-equiprobable sources is,

$$\bar{R}(n) = H - \frac{1}{2n} \log_2 n + O\left(\frac{1}{n}\right), \quad (30)$$

which is also refined to show that, if $\iota_X(X)$ is non-lattice,² then,

$$\bar{R}(n) = H - \frac{1}{2n} \log_2(8\pi e \sigma^2 n) + o\left(\frac{1}{n}\right), \quad (31)$$

where,

$$\sigma^2 = \text{Var}(\iota_X(X)), \quad (32)$$

is the *varentropy* or *minimal coding variance* [11] of P_X . In contrast, when a prefix-free condition is imposed, we have the well-known behavior (see, e.g., [4]),

$$\bar{R}_p(n) = H + O\left(\frac{1}{n}\right), \quad (33)$$

for any source for which $H = \lim_{n \rightarrow \infty} \frac{1}{n} H(X^n)$ exists.

¹A legacy of the Kraft inequality mindset, the term “ideal codelength” is sometimes used for $\iota_X(X)$. This is inappropriate in view of the fact that the optimum codelengths are in fact *bounded above* by $\iota_X(X)$; see Section II. Therefore, these “ideal codelengths” are neither ideal nor are they actual codelengths.

²A discrete random variable is *lattice* if all its masses are on a subset of some lattice $\{\nu + k\varsigma; k \in \mathbb{Z}\}$.

For a non-equiprobable source such that $\iota_X(X)$ is non-lattice, Strassen [26] claims³ the following Gaussian approximation result as a refinement of (26):

$$\begin{aligned} R^*(n, \epsilon) = & H + \frac{\sigma}{\sqrt{n}} Q^{-1}(\epsilon) - \frac{1}{2n} \log_2 \left(2\pi\sigma^2 n e^{(Q^{-1}(\epsilon))^2} \right) \\ & + \frac{\mu_3}{6\sigma^2 n} ((Q^{-1}(\epsilon))^2 - 1) + o\left(\frac{1}{n}\right). \end{aligned} \quad (34)$$

Here, $Q(x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-t^2/2} dt$ denotes the standard Gaussian tail function, σ^2 is the varentropy of P_X defined in (32), and μ_3 is the third centered absolute moment of the information random variable (29).

Kontoyiannis [11] gives a different kind of Gaussian approximation for the codelengths $\ell(f_n(X^n))$ of arbitrary prefix codes f_n on memoryless data X^n , showing that, with probability one, $\ell(f_n(X^n))$ is eventually lower bounded by a random variable that has an approximately Gaussian distribution,

$$\ell(f_n(X^n)) \geq Z_n \text{ where } Z_n \stackrel{\mathcal{D}}{\approx} N(nH, n\sigma^2); \quad (35)$$

and σ^2 is the varentropy as in (32). Therefore, the codelengths $\ell(f_n(X^n))$ will have at least Gaussian fluctuations of $O(\sqrt{n})$; this is further sharpened in [11] to a corresponding law of the iterated logarithm, stating that, with probability one, the compressed lengths $\ell(f_n(X^n))$ will have fluctuations of $O(\sqrt{n \ln \ln n})$, infinitely often: with probability one,

$$\limsup_{n \rightarrow \infty} \frac{\ell(f_n(X^n)) - H(X^n)}{\sqrt{2n \ln \ln n}} \geq \sigma. \quad (36)$$

Both results (35) and (36) are shown to hold for Markov sources as well as for a wide class of mixing sources with infinite memory.

E. Outline of main new results

Section II gives a general analysis of the distribution of the lengths of an optimal lossless code for any discrete information source, which may or may not produce fixed-length strings of symbols. First, in Theorems 2 and 3 we give simple achievability and converse bounds, showing that the distribution function of the optimal codelengths, $\mathbb{P}[\ell(f^*(X)) \leq t]$, is intimately related to the distribution of the information random variable, $\mathbb{P}[\iota_X(X) \leq t]$. Also we observe that the optimal codelengths $\ell(f^*(X))$ are always bounded above by $\iota_X(X)$, but Theorem 4 states that they cannot be significantly smaller than $\iota_X(X)$ with high probability. The corresponding result for prefix codes, originally derived in [2], [11], is stated in Theorem 5.

Theorem 6 offers an exact, non-asymptotic expression for best achievable rate $R^*(n, \epsilon)$. So far, no other problem in information theory has yielded an exact non-asymptotic formula for the fundamental limit. An exact expression for the average probability of error achieved by (almost-lossless) random binning, is given in Theorem 7.

General non-asymptotic and asymptotic results for the expected optimal length, $\bar{R}(n) = (1/n)\mathbb{E}[\ell(f_n^*(X^n))]$, are obtained in Section III. Attained by the Huffman code, the minimal

³See the discussion in Section V regarding Strassen's claimed proof of this result.

average length of prefix codes is unknown. However, dropping the extraneous prefix constraint for non-symbol-by-symbol codes results in an explicit formula for the minimal average length.

In Section IV we revisit the refined asymptotic results (35) and (36) of [11], and show that they remain valid for general (not necessarily prefix) compressors, and for a broad class of possibly infinite-memory sources.

Section V examines in detail the finite-blocklength behavior of the fundamental limit $R^*(n, \epsilon)$ for the case of memoryless sources. We prove tight, non-asymptotic and easily computable bounds for $R^*(n, \epsilon)$; specifically, combining the results of Theorems 16 and 17 implies the following approximation for finite blocklengths n :

Gaussian approximation I: For every memoryless source, the best achievable rate $R^*(n, \epsilon)$ satisfies:

$$nR^*(n, \epsilon) \approx nH + \sigma\sqrt{n}Q^{-1}(\epsilon) - \frac{1}{2}\log_2 n, \quad (37)$$

where the approximation is accurate up to $O(1)$ terms; the same holds for $R_p(n, \epsilon)$ in the case of prefix codes.

The approximation (37) is established by combining the general results of Section II with the classical Berry-Esséen bound [15], [21]. This approximation is made precise in a non-asymptotic way, and all the constants involved are explicitly identified.

In Section VI, achievability and converse bounds (Theorems 18 and 19) are established for $R^*(n, \epsilon)$, in the case of general ergodic Markov sources. Those results are analogous (though slightly weaker) to those in Section V.

We also define the varentropy rate of an arbitrary source as the limiting normalized variance of the information random variables $\iota_{X^n}(X^n)$, and we show that, for Markov chains, it plays the same role as the varentropy defined in (32) for memoryless sources. Those results in particular imply the following:

Gaussian approximation II: For any ergodic Markov source with entropy rate H and varentropy rate σ^2 , the blocklength $n^*(R, \epsilon)$ required for the compression rate to exceed $(1 + \eta)H$ with probability no greater than $\epsilon > 0$, satisfies,

$$n^*((1 + \eta)H, \epsilon) \approx \frac{\sigma^2}{H^2} \left(\frac{Q^{-1}(\epsilon)}{1 + \eta} \right)^2. \quad (38)$$

[See Section IV for the general definition of the varentropy rate σ^2 , and the discussion in Section VI for details.]

Finally, Section VII defines the source dispersion D as the limiting normalized variance of the optimal codelengths. In effect, the dispersion gauges the time one must wait for the source realization to become typical within a given probability, as in (38) above, with D in place of σ^2 . For a large class of sources (including ergodic Markov chains of any order), the dispersion D is shown to equal the varentropy rate σ^2 of the source.

II. NON-ASYMPTOTIC BOUNDS FOR ARBITRARY SOURCES

In this section we analyze the best achievable compression performance on a completely general discrete random source. In particular, (except where noted) we do not necessarily assume that the alphabet is finite and we do not exploit the fact that in the original problem we are interested in compressing a block of n symbols. In this way we even encompass the case where the source string length is a priori unknown at the decompressor. Thus, we consider a given probability mass function P_X defined on an arbitrary finite alphabet \mathcal{X} , which may (but is not assumed to) consist of variable-length strings drawn from some alphabet. The results can then be particularized to the setting in Section I, letting $\mathcal{X} \leftarrow \mathcal{A}^n$ and $P_X \leftarrow P_{X^n}$. Conversely, we can simply let $n = 1$ in Section I to yield the setting in this section.

The best achievable rate $R^*(n, \epsilon)$ at blocklength $n = 1$ is abbreviated as $R^*(\epsilon) = R^*(1, \epsilon)$. By definition, $R^*(\epsilon)$ is the lowest R such that,

$$\mathbb{P}[\ell(f^*(X)) > R] \leq \epsilon, \quad (39)$$

which is equal to the quantile function⁴ of the integer-valued random variable $\ell(f^*(X))$ evaluated at $1 - \epsilon$.

A. Achievability bound

Recall the definition of the information random variable $\iota_X(X)$ in (29). Our goal is to express the distribution of the optimal codelengths $\ell(f^*(X))$ in terms of the distribution of $\iota_X(X)$. The first such result is the following simple and powerful upper bound (e.g. [30]) on the tail of the distribution of the minimum rate.

Theorem 2: For any $a \geq 0$,

$$\mathbb{P}[\ell(f^*(X)) \geq a] \leq \mathbb{P}[\iota_X(X) \geq a]. \quad (40)$$

Proof: Since the labeling of the values taken by the random variable X is immaterial, it simplifies notation in the proofs to assume that the elements of \mathcal{X} are integer-valued with decreasing probabilities: $P_X(1) \geq P_X(2) \geq \dots$. Then, for all $i = 1, 2, \dots$ we have the fundamental relationships:

$$\ell(f^*(i)) = \lfloor \log_2 i \rfloor \quad (41)$$

$$P_X(i) \leq \frac{1}{i}. \quad (42)$$

Therefore,

$$\mathbb{P}[\ell(f^*(X)) \geq a] = \mathbb{P}[\lfloor \log_2 X \rfloor \geq a] \quad (43)$$

$$\leq \mathbb{P}[\log_2 X \geq a] \quad (44)$$

$$\leq \mathbb{P}[\iota_X(X) \geq a], \quad (45)$$

where (45) follows from (42). ■

⁴The quantile function $\mathcal{Q}: [0, 1] \rightarrow \mathbb{R}$ is the “inverse” of the cumulative distribution function F . Specifically, $\mathcal{Q}(\alpha) = \min\{x: F(x) = \alpha\}$ if the set is nonempty; otherwise α lies within a jump $\lim_{x \uparrow x_\alpha} F(x) < \alpha < F(x_\alpha)$ and we define $\mathcal{Q}(\alpha) = x_\alpha$.

Before moving on, we point out that at the core of the above proof is a simple but crucial observation: not only does the distribution function of the optimal codelengths $\ell(f^*(X))$ dominate that of $\iota_X(X)$, but we in fact *always* have,

$$\ell(f^*(x)) \leq \iota_X(x), \text{ for all } x \in \mathcal{X}. \quad (46)$$

This will be used repeatedly, throughout the rest of the paper. Also, a simple inspection of the proof shows that Theorem 2 as well as (46) remain valid even in the case of sources X with a countably infinite alphabet.

Theorem 2 is the starting point for the achievability result for $R^*(n, \epsilon)$ established for Markov sources in Theorem 18.

B. Converse bounds

In Theorem 3 we give a corresponding converse result; cf. [30]. It will be used later to obtain sharp converse bounds for $R^*(n, \epsilon)$ for memoryless and Markov sources, in Theorems 17 and 19, respectively.

Theorem 3: For any nonnegative integer k ,

$$\max_{\tau > 0} \{ \mathbb{P}[\iota_X(X) \geq k + \tau] - 2^{-\tau} \} \leq \mathbb{P}[\ell(f^*(X)) \geq k]. \quad (47)$$

Proof: As in the proof of Theorem 2, we label the values taken by X as the positive integers in decreasing probabilities. Fix an arbitrary $\tau > 0$. Define:

$$\mathcal{L} = \{i \in \mathcal{X} : P_X(i) \leq 2^{-k-\tau}\} \quad (48)$$

$$\mathcal{C} = \{1, 2, \dots, 2^k - 1\}. \quad (49)$$

Then, abbreviating $P_X(\mathcal{B}) = \mathbb{P}[X \in \mathcal{B}] = \sum_{i \in \mathcal{B}} P_X(i)$, for any $\mathcal{B} \subset \mathcal{X}$,

$$\mathbb{P}[\iota_X(X) \geq k + \tau] = P_X(\mathcal{L}) \quad (50)$$

$$= P_X(\mathcal{L} \cap \mathcal{C}) + P_X(\mathcal{L} \cap \mathcal{C}^c) \quad (51)$$

$$\leq P_X(\mathcal{L} \cap \mathcal{C}) + P_X(\mathcal{C}^c) \quad (52)$$

$$\leq (2^k - 1)2^{-k-\tau} + P_X(\mathcal{C}^c) \quad (53)$$

$$< 2^{-\tau} + \mathbb{P}[\lfloor \log_2 X \rfloor \geq k] \quad (54)$$

$$= 2^{-\tau} + \mathbb{P}[\ell(f^*(X)) \geq k], \quad (55)$$

where (55) follows in view of (41). ■

Next we give another general converse bound, similar to that of Theorem 3, where this time we directly compare the codelengths $\ell(f(X))$ of an arbitrary compressor with the values of the information random variable $\iota_X(X)$. Whereas from (46) we know that $\ell(f(X))$ is always smaller than $\iota_X(X)$, Theorem 4 says that it cannot be much smaller with high probability. This is a natural analog of the corresponding converse established for prefix compressors in [2], and stated as Theorem 5 below.

Applying to a finite-alphabet source, Theorem 4 is the key bound in the derivation of all the pointwise asymptotic results of Section IV, Theorems 11, 12 and 13. It is also the main technical ingredient of the proof of Theorem 22 in Section VII stating that the source dispersion is equal to its varentropy.

Theorem 4: For any compressor f and any $\tau > 0$,

$$\mathbb{P}[\ell(f(X)) \leq \iota_X(X) - \tau] \leq 2^{-\tau} (\lceil \log_2 |\mathcal{X}| \rceil + 1) \quad (56)$$

Proof: Letting $\mathbb{I}\{A\}$ denote the indicator function of the event A , the probability in (56) can be bounded by

$$\mathbb{P}[\ell(f(X)) \leq \iota_X(X) - \tau] = \sum_{x \in \mathcal{X}} P_X(x) \mathbb{I}\{P_X(x) \leq 2^{-\tau - \ell(f(x))}\} \quad (57)$$

$$\leq 2^{-\tau} \sum_{x \in \mathcal{X}} 2^{-\ell(f(x))}, \quad (58)$$

$$\leq 2^{-\tau} \sum_{j=0}^{\lceil \log_2 |\mathcal{X}| \rceil} 2^j 2^{-j} \quad (59)$$

where the sum in (58) is maximized if f assigns a string of length $j+1$ only if it also assigns all strings of length j . Therefore, (59) holds because that code contains all the strings of lengths $0, 1, \dots, \lceil \log_2 |\mathcal{X}| \rceil - 1$ plus $|\mathcal{X}| - 2^{\lceil \log_2 |\mathcal{X}| \rceil} + 1 \leq 2^{\lceil \log_2 |\mathcal{X}| \rceil}$ strings of length $\lceil \log_2 |\mathcal{X}| \rceil$. ■

We saw in Theorem 1 that the optimum prefix code under the criterion of minimum excess length probability incurs a penalty of at most one bit. The following elementary converse is derived in [2], [11]; its short proof is included for completeness. Indeed, the statements and proofs of Theorems 4 and 5 are close parallels.

Theorem 5: For any prefix code f , and any $\tau \geq 0$:

$$\mathbb{P}[\ell(f(X)) < \iota_X(X) - \tau] \leq 2^{-\tau}. \quad (60)$$

Proof: We have, as in the proof of Theorem 4 leading to (58),

$$\mathbb{P}[\ell(f(X)) < \iota_X(X) - \tau] < 2^{-\tau} \sum_{x \in \mathcal{X}} 2^{-\ell(f(x))} \quad (61)$$

$$\leq 2^{-\tau}, \quad (62)$$

where (62) is Kraft's inequality. ■

C. Exact fundamental limit

The following result expresses the non-asymptotic data compression fundamental limit $R^*(\epsilon) = R^*(1, \epsilon)$ as a function of the source information spectrum.

Theorem 6: For all $a \geq 0$, the exact minimum rate compatible with given excess-length probability satisfies,

$$R^*(\epsilon) = \lceil \log_2 (1 + M(2^a)) \rceil - 1, \quad (63)$$

with,

$$\epsilon = \mathbb{P}[\iota_X(X) \geq a], \quad (64)$$

where $M(\beta)$ denotes the number of masses with probability strictly larger than $\frac{1}{\beta}$, and which can be expressed as:

$$M(\beta) = \beta \mathbb{P}[\iota_X(X) < \log_2 \beta] - \int_1^\beta \mathbb{P}[\iota_X(X) \leq \log_2 t] dt. \quad (65)$$

Proof: As above, the values taken by X are labeled as the positive integers in order of decreasing probability. By the definition of $M(\cdot)$, for any positive integer i , and $a > 0$,

$$P_X(i) \leq 2^{-a} \iff \log_2(1 + M(2^a)) \leq \log_2 i, \quad (66)$$

and it is easy to check that:

$$\lceil \alpha \rceil - 1 < \lfloor \log_2 i \rfloor \iff \alpha \leq \log_2 i. \quad (67)$$

Therefore, letting $\alpha = \log_2(1 + M(2^a))$ and letting the integer-valued X take the role of i , we obtain that (39) is satisfied with equality if R is given by the right side of (63). Any smaller value of R would prevent (39) from being satisfied.

The proof of (65) follows a sequence of elementary steps:

$$M(\beta) = \sum_{x \in \mathcal{X}} \mathbb{I} \left\{ P_X(x) > \frac{1}{\beta} \right\} \quad (68)$$

$$= \mathbb{E} \left[\frac{\mathbb{I} \{ P_X(X) > \frac{1}{\beta} \}}{P_X(X)} \right] \quad (69)$$

$$= \int_0^\infty \mathbb{P} \left[\frac{\mathbb{I} \{ P_X(X) > \frac{1}{\beta} \}}{P_X(X)} > t \right] dt \quad (70)$$

$$= \int_0^\beta \mathbb{P} \left[\frac{1}{\beta} < P_X(X) < \frac{1}{t} \right] dt \quad (71)$$

$$= \int_0^\beta \mathbb{P} \left[\frac{1}{\beta} < P_X(X) \right] - \mathbb{P} \left[P_X(X) \geq \frac{1}{t} \right] dt \quad (72)$$

$$= \beta \mathbb{P}[\iota_X(X) < \log_2 \beta] - \int_1^\beta \mathbb{P}[\iota_X(X) \leq \log_2 t] dt. \quad (73)$$

■

While Theorem 6 gives $R^*(\epsilon) = R^*(1, \epsilon)$ exactly for those ϵ which correspond to values taken by the complementary cumulative distribution function of the information random variable $\iota_X(X)$, a continuous sweep of $a > 0$ gives a very dense grid of values, unless X (whose alphabet size typically grows exponentially with n in the fixed-to-variable setup) takes values in a very small alphabet. From the value of a we can obtain the probability in the right side of (64). The optimum code achieves that excess probability $\epsilon = \mathbb{P}[\ell(f^*(X)) \geq \ell_a]$ for lengths equal to,

$$\ell_a = \lceil a + \log_2(2^{-a} + 2^{-a} M(2^a)) \rceil, \quad (74)$$

where the second term is negative and represents the exact gain with respect to the information spectrum of the source.

For later use we observe that, if we let $M_X^+(\beta)$ be the number of masses with probability larger or equal than $\frac{1}{\beta}$, then,⁵

$$M_X^+(\beta) = \sum_{x \in \mathcal{X}} \mathbb{I} \left\{ P_X(x) \geq \frac{1}{\beta} \right\} \quad (75)$$

$$= \mathbb{E} [\exp(\iota_X(X)) \mathbb{I} \{ \iota_X(X) \leq \log_2 \beta \}]. \quad (76)$$

Figure 1 shows the cumulative distribution functions of $\ell(f^*(X))$ and $\iota_X(X)$ when X is a binomially distributed random variable: the number of tails obtained in 10,000 fair coin flips. Therefore, $\iota_X(X)$ ranges from $6.97 \approx 10,000 - \log_2 \binom{10000}{5000}$ to 10,000 and,

$$H(X) = 7.69 \quad (77)$$

$$\mathbb{E}[\ell(f^*(X))] = 6.29, \quad (78)$$

where all figures are in bits.

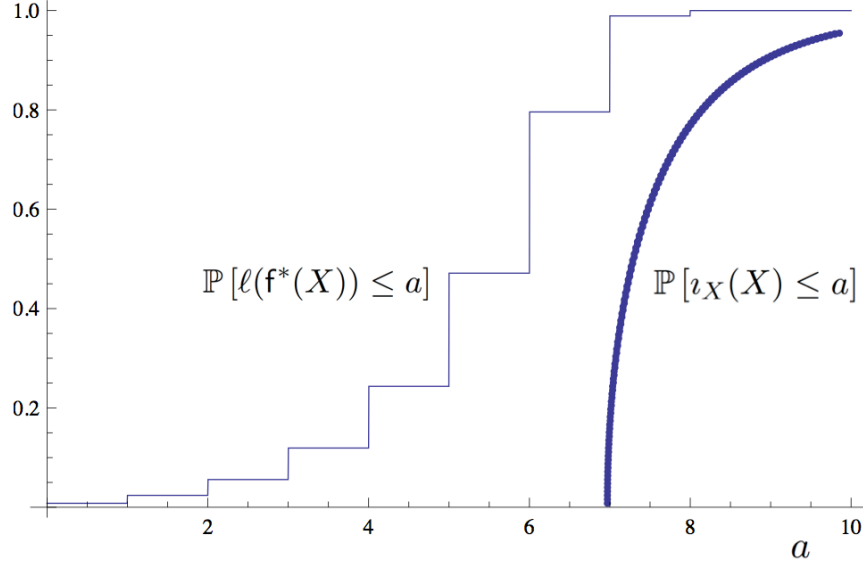


Fig. 1: Cumulative distribution functions of $\ell(f^*(X))$ and $\iota_X(X)$ when X is the number of tails obtained in 10,000 fair coin flips.

D. Exact behavior of random binning

The following result gives an exact expression for the performance of random binning for arbitrary sources, as a function of the cumulative distribution function of the random variable $\iota_X(X)$ via (65). In binning, the compressor is no longer constrained to be an injective mapping. When the label received by the decompressor can be explained by more than one source realization, it chooses the most likely one, breaking ties arbitrarily. (Cf. [23] for the exact performance of random coding in channel coding.)

⁵Where typographically convenient we use $\exp(a) = 2^a$.

Theorem 7: Averaging uniformly over all binning compressors $f: \mathcal{X} \rightarrow \{1, 2, \dots, N\}$, results in an expected error probability equal to,

$$1 - \mathbb{E} \left[\sum_{\ell=0}^{J(X)-1} \frac{\binom{J(X)-1}{\ell}}{N^\ell(1+\ell)} \left(1 - \frac{1}{N}\right)^{M(\frac{1}{P_X(X)})+J(X)-\ell-1} \right], \quad (79)$$

where $M(\cdot)$ is given in (65), and the number of masses whose probability is equal to $P_X(x)$ is denoted by:

$$J(x) = \frac{\mathbb{P}[P_X(X) = P_X(x)]}{P_X(x)}. \quad (80)$$

Proof: For the purposes of the proof, it is convenient to assume that ties are broken uniformly at random among the most likely source outcomes in the bin. To verify (79), note that, given that the source realization is x_0 :

- 1) The number of masses with probability strictly higher than that of x_0 is $M(\frac{1}{P_X(x_0)})$;
- 2) Correct decompression of x_0 requires that any x with $P_X(x) > P_X(x_0)$ not be assigned to the same bin as x_0 . This occurs with probability:

$$\left(1 - \frac{1}{N}\right)^{M(\frac{1}{P_X(x_0)})}; \quad (81)$$

- 3) If there are ℓ masses with the same probability as x_0 in the same bin, correct decompression occurs with probability $\frac{1}{1+\ell}$.
- 4) The probability that there are ℓ masses with the same probability as x_0 in the same bin is equal to:

$$\binom{J(x_0)-1}{\ell} \left(1 - \frac{1}{N}\right)^{J(x_0)-\ell-1} \frac{1}{N^\ell}. \quad (82)$$

Then, (79) follows since all the bin assignments are independent. ■

Theorem 7 leads to an achievability bound for both almost-lossless fixed-to-fixed compression and lossless fixed-to-variable compression. However, in view of the simplicity and tightness of Theorem 2, the main usefulness of Theorem 7 is to gauge the suboptimality of random binning in the finite (in fact, rather short because of computational complexity) blocklength regime.

III. MINIMAL EXPECTED LENGTH

Recall the definition of the best achievable rate $\bar{R}(n)$ in Section I, expressed in terms of f_n^* as in (8). An immediate consequence of Theorem 2 is the bound,

$$n\bar{R}(n) = \mathbb{E}[\ell(f^*(X))] \leq H(X), \quad (83)$$

which goes back at least to the work of Wyner [32]. Indeed, by lifting the prefix condition it is possible to beat the entropy on average as we saw in the asymptotic results (30) and (31). Lower bounds on the minimal average length as a function of $H(X)$ can be found in [27] and references therein. An explicit expression can be obtained easily by labeling the outcomes as the positive integers with decreasing probabilities as in the proof of Theorem 2:

$$\mathbb{E}[\ell(f^*(X))] = \mathbb{E}[\lfloor \log_2 X \rfloor] \quad (84)$$

$$= \sum_{k=1}^{\infty} \mathbb{P}[\lfloor \log_2 X \rfloor \geq k] \quad (85)$$

$$= \sum_{k=1}^{\infty} \mathbb{P}[X \geq 2^k]. \quad (86)$$

Example 2: The average number of bits required to encode at which flip of a fair coin the first tail appears is equal to,

$$\sum_{k=1}^{\infty} \mathbb{P}[X \geq 2^k] = \sum_{k=1}^{\infty} \sum_{j=2^k}^{\infty} 2^{-j} \quad (87)$$

$$= 2 \sum_{k=1}^{\infty} 2^{2^{-k}} \quad (88)$$

$$\approx 0.632843, \quad (89)$$

since, in this case, X is a geometric random variable with $\mathbb{P}[X = j] = 2^{-j}$. In contrast, imposing a prefix constraint disables any compression: the optimal prefix code consists of all, possibly empty, strings of 0s terminated by 1, achieving an average length of 2.

Example 3: If X_M is equiprobable on a set of M elements, then:

1)

$$\mathbb{E}[\ell(f^*(X_M))] = \lfloor \log_2 M \rfloor + \frac{1}{M} (2 + \lfloor \log_2 M \rfloor - 2^{\lfloor \log_2 M \rfloor + 1}), \quad (90)$$

which simplifies to,

$$\mathbb{E}[\ell(f^*(X_M))] = \frac{(M+1)\log_2(M+1)}{M} - 2, \quad (91)$$

when $M+1$ is a power of 2.

2)

$$\limsup_{M \rightarrow \infty} H(X_M) - \mathbb{E}[\ell(f^*(X_M))] = 2 \quad (92)$$

$$\liminf_{M \rightarrow \infty} H(X_M) - \mathbb{E}[\ell(f^*(X_M))] = 1 + \log_2 e - \log_2 \log_2 e, \quad (93)$$

where the entropy is expressed in bits.

Theorem 8: For any source $\mathbf{X} = \{P_{X^n}\}_{n=1}^\infty$ with finite entropy rate,

$$H(\mathbf{X}) = \limsup_{n \rightarrow \infty} \frac{1}{n} H(X^n) < \infty, \quad (94)$$

the normalized minimal average length satisfies:

$$\limsup_{n \rightarrow \infty} \bar{R}(n) = H(\mathbf{X}). \quad (95)$$

Proof: The achievability (upper) bound in (95) holds in view of (83). In the reverse direction, we invoke the bound [1]:

$$H(X^n) - \mathbb{E}[\ell(f_n^*(X^n))] \leq \log_2(H(X^n) + 1) + \log_2 e. \quad (96)$$

Upon dividing both sides of (96) by n and taking \limsup the desired result follows, since for any $\delta > 0$, for all sufficiently large n , $H(X^n) \leq nH(\mathbf{X}) + n\delta$. ■

In view of (33), we see that the penalty incurred on the *average rate* by the prefix condition vanishes asymptotically in the very wide generality allowed by Theorem 8. In fact, the same proof we used for Theorem 8 shows the following result:

Theorem 9: For any (not necessarily serial) source $\mathbf{X} = \{P_{X^{(n)}}\}_{n=1}^\infty$,

$$\lim_{n \rightarrow \infty} \frac{\bar{R}(n)}{\frac{1}{n} H(X^{(n)})} = \lim_{n \rightarrow \infty} \frac{\mathbb{E}[\ell(f_n^*(X^{(n)}))]}{H(X^{(n)})} = 1, \quad (97)$$

as long as $H(X^{(n)})$ diverges, where $X^{(n)} \in \mathcal{A}_n$, an alphabet which is not necessarily a Cartesian product.

IV. POINTWISE ASYMPTOTICS

A. Normalized pointwise redundancy

Before turning to the precise evaluation of the best achievable rate $R^*(n, \epsilon)$, in this section we examine the asymptotic behavior of the normalized difference between the codelength and the information (sometimes known as the pointwise redundancy).

Theorem 10: For any discrete source and any divergent deterministic sequence κ_n such that,

$$\lim_{n \rightarrow \infty} \frac{\log n}{\kappa_n} = 0, \quad (98)$$

we have:

(a) For any sequence $\{f_n\}$ of codes:

$$\liminf_{n \rightarrow \infty} \frac{1}{\kappa_n} (\ell(f_n(X^n)) - \iota_{X^n}(X^n)) \geq 0, \quad \text{w.p.1.} \quad (99)$$

(b) The sequence of optimal codes $\{f_n^*\}$ achieves:

$$\liminf_{n \rightarrow \infty} \frac{1}{\kappa_n} (\ell(f_n^*(X^n)) - \iota_{X^n}(X^n)) = 0, \quad \text{w.p.1.} \quad (100)$$

Proof: (a) We invoke the general converse in Theorem 4, with X^n and \mathcal{A}^n in place of X and \mathcal{X} , respectively. Fixing arbitrary $\epsilon > 0$ and letting $\tau = \tau_n = \epsilon \kappa_n$, we obtain that,

$$\mathbb{P}[\ell(f_n(X^n)) \leq \iota_{X^n}(X^n) - \epsilon \kappa_n] \leq 2^{\log_2 n - \epsilon \kappa_n} (\log_2 |\mathcal{A}| + 1) \quad (101)$$

which is summable in n . Therefore, the Borel-Cantelli lemma implies that the \limsup of the event on the left side of (101) has zero probability, or equivalently, with probability one,

$$\ell(f_n(X^n)) - \iota_{X^n}(X^n) \geq -\epsilon \kappa_n$$

is violated only a finite number of times. Since ϵ can be chosen arbitrarily small, (99) follows.

Part (b) follows from (a) and (46). \blacksquare

B. Stationary Ergodic Sources

Theorem 8 states that for any discrete process \mathbf{X} the expected rate of the optimal codes f_n^* satisfy,

$$\limsup_{n \rightarrow \infty} \frac{1}{n} \mathbb{E}[\ell(f_n^*(X^n))] = H(\mathbf{X}). \quad (102)$$

The next result shows that if the source is stationary and ergodic, then the same asymptotic relation holds not just in expectation, but also with probability 1. Moreover, no compressor can beat the entropy rate asymptotically with positive probability. The corresponding results for prefix codes were established in [2], [10], [11].

Theorem 11: Suppose that $\{X_n\}$ is a stationary ergodic source with entropy rate H .

(i) For any sequence $\{f_n\}$ of codes,

$$\liminf_{n \rightarrow \infty} \frac{1}{n} \ell(f_n(X^n)) \geq H, \quad \text{w.p.1.} \quad (103)$$

(ii) The sequence of optimal codes $\{f_n^*\}$ achieves,

$$\lim_{n \rightarrow \infty} \frac{1}{n} \ell(f_n^*(X^n)) = H, \quad \text{w.p.1.} \quad (104)$$

Proof: The Shannon-Macmillan-Breiman theorem states that,

$$\frac{1}{n} \iota_{X^n}(X^n) \rightarrow H, \quad \text{w.p.1.} \quad (105)$$

Therefore, the result is an immediate consequence of Theorem 10 with $\kappa_n = n$. \blacksquare

C. Stationary Ergodic Markov Sources

We assume that the source is a stationary ergodic (first-order) Markov chain, with transition kernel,

$$P_{X'|X}(x' | x) \quad (x, x') \in \mathcal{A}^2, \quad (106)$$

on the finite alphabet \mathcal{A} . Further restricting the source to be Markov enables us to analyze more precisely the behavior of the information random variables and, in particular, we will show that the zero-mean random variables,

$$Z_n = \frac{\iota_{X^n}(X^n) - H(X^n)}{\sqrt{n}}, \quad (107)$$

are asymptotically normal with variance given by the varentropy rate, which generalizes the notion in (32).

Definition 1: The varentropy rate of a random process $\mathbf{X} = \{P_{X^n}\}_{n=1}^\infty$ is

$$\sigma^2 = \limsup_{n \rightarrow \infty} \frac{1}{n} \text{Var}(\iota_{X^n}(X^n)). \quad (108)$$

Some remarks are in order:

- If \mathbf{X} is a stationary memoryless process each of whose letters is distributed according to P_X , then the varentropy rate of \mathbf{X} is equal to the varentropy of X . The varentropy of X is zero if and only if it is equiprobable on its support.
- In contrast to the first moment, we do not know whether stationarity is sufficient for $\limsup = \liminf$ in (108).
- While the entropy-rate of a Markov chain admits a two-letter expression, the varentropy does not. In particular, if $\sigma^2(a)$ denotes the varentropy of the distribution $P_{X'|X}(\cdot | a)$, then the varentropy of the chain is, in general, not given by $\mathbb{E}[\sigma^2(X_0)]$.
- The varentropy rate of Markov sources is typically nonzero. For example, for a first order Markov chain it was observed in [33], [12] that $\sigma^2 = 0$ if and only if the source satisfies the following *deterministic equipartition property*: Every string x^{n+1} that starts and ends with the same symbol, has probability (given that $X_1 = x_1$) q^n , for some constant $0 \leq q \leq 1$.

Theorem 12: Let $\{X_n\}$ be a stationary ergodic finite-state Markov chain.

- (i) The varentropy rate σ^2 is also equal to the corresponding \liminf of the normalized variances in (108), and it is finite.

- (ii) The normalized information random variables are asymptotically normal, in the sense that, as $n \rightarrow \infty$,

$$\frac{\iota_{X^n}(X^n) - H(X^n)}{\sqrt{n}} \longrightarrow N(0, \sigma^2), \quad (109)$$

in distribution.

- (iii) The normalized information random variables satisfy a corresponding law of the iterated logarithm:

$$\limsup_{n \rightarrow \infty} \frac{\iota_{X^n}(X^n) - H(X^n)}{\sqrt{2n \ln \ln n}} = \sigma, \quad \text{w.p.1} \quad (110)$$

$$\liminf_{n \rightarrow \infty} \frac{\iota_{X^n}(X^n) - H(X^n)}{\sqrt{2n \ln \ln n}} = -\sigma, \quad \text{w.p.1} \quad (111)$$

Proof:

- (i) and (ii): Consider the bivariate Markov chain $\{\tilde{X}_n = (X_n, X_{n+1})\}$ on the alphabet $\mathcal{B} = \{(x, y) \in \mathcal{A}^2 : P_{X'|X}(y|x) > 0\}$ and the function $f: \mathcal{B} \rightarrow \mathbb{R}$ defined by

$$f(x, y) = \iota_{X'|X}(y|x). \quad (112)$$

Since $\{X_n\}$ is stationary and ergodic, so is $\{\tilde{X}_n\}$, hence, by the central limit theorem for functions of Markov chains [5]

$$\frac{1}{\sqrt{n}} (\iota_{X^n|X_1}(X^n|X_1) - H(X^n|X_1)) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n-1} (f(\tilde{X}_i) - \mathbb{E}[f(\tilde{X}_i)]) \quad (113)$$

converges in distribution to the zero-mean Gaussian law with finite variance

$$\Sigma^2 = \lim_{n \rightarrow \infty} \frac{1}{n} \text{Var}(\iota_{X^n|X_1}(X^n|X_1)). \quad (114)$$

Furthermore, since

$$\iota_{X^n}(X^n) - H(X^n) = \iota_{X^n|X_1}(X^n|X_1) - H(X^n|X_1) + (\iota_{X_1}(X_1) - H(X_1)) \quad (115)$$

where the second term is bounded, (109) must hold and we must have $\Sigma^2 = \sigma^2$.

- (iii) Normalizing (113) by $\sqrt{2n \ln \ln n}$ in lieu of \sqrt{n} , we can invoke the law of the iterated logarithm for functions of Markov chains [5] to show that the \limsup / \liminf of the sum behave as claimed. Since upon normalization, the second term in the right side of (115), vanishes almost surely, $\iota_{X^n}(X^n) - H(X^n)$ must satisfy the same behavior. ■

Together with Theorem 10 particularized to $\kappa_n = \sqrt{n}$, we conclude that the normalized deviation of the optimal codelengths from the entropy rate satisfies

$$\frac{\ell(\mathbf{f}_n^*(X^n)) - H}{\sqrt{n}} \longrightarrow N(0, \sigma^2) \quad (116)$$

which is the same behavior as that exhibited by the Shannon prefix code [11], so as far as the pointwise \sqrt{n} asymptotics the prefix constraint does not entail loss of efficiency. Similarly, the following result readily follows from Theorem 12 and Theorem 10 with $\kappa_n = \sqrt{2n \ln \ln n}$.

Theorem 13: Suppose $\{X_n\}$ is a stationary ergodic Markov chain with entropy rate H and varentropy rate σ^2 . Then:

(i) For any sequence of codes $\{f_n\}$:

$$\limsup_{n \rightarrow \infty} \frac{\ell(f_n(X^n)) - H(X^n)}{\sqrt{2n \ln \ln n}} \geq \sigma, \quad \text{w.p.1;} \quad (117)$$

$$\liminf_{n \rightarrow \infty} \frac{\ell(f_n(X^n)) - H(X^n)}{\sqrt{2n \ln \ln n}} \geq -\sigma, \quad \text{w.p.1.} \quad (118)$$

(ii) The sequence of optimal codes $\{f_n^*\}$ achieves the bounds in (117) and (118) with equality.

The Markov sufficient condition in Theorem 12 enabled the application of the central limit theorem and the law of the iterated logarithm to the sum in (113). According to Theorem 9.1 of [22] a more general sufficient condition is that $\{X_n\}$ be a stationary process with $\alpha(d) = O(d^{-336})$ and $\gamma(d) = O(d^{-48})$, with the mixing coefficients:

$$\gamma(d) = \max_{a \in \mathcal{A}} \mathbb{E} \left| \imath_{X_0|X_{-\infty}^{-1}}(a|X_{-1}, X_{-2}, \dots) - \imath_{X_0|X_{-d}^{-1}}(a|X_{-1}, X_{-2}, \dots, X_{-d}) \right| \quad (119)$$

$$\alpha(d) = \sup \left\{ |\mathbb{P}(B \cap A) - \mathbb{P}(B)\mathbb{P}(A)| ; A \in \mathcal{F}_{-\infty}^0, B \in \mathcal{F}_d^\infty \right\}. \quad (120)$$

Here $\mathcal{F}_{-\infty}^0$ and \mathcal{F}_d^∞ denote the σ -algebras generated by the collections of random variables (X_0, X_{-1}, \dots) and (X_d, X_{d+1}, \dots) , respectively. The $\alpha(d)$ are the *strong mixing* coefficients [3] of $\{X_n\}$, and the $\gamma(d)$ were introduced by Ibragimov in [8]. Although these mixing conditions may be hard to verify in practice, they are fairly weak in that they require only polynomial decay of $\alpha(d)$ and $\gamma(d)$. In particular, any ergodic Markov chain of any order satisfies these conditions.

V. GAUSSIAN APPROXIMATION FOR MEMORYLESS SOURCES

We turn our attention to the non-asymptotic behavior of the best rate $R^*(n, \epsilon)$ that can be achieved when compressing a stationary memoryless finite-alphabet source $\{X_n \in \mathcal{A}\}$ with marginal distribution P_X , whose entropy and varentropy are denoted by H and σ^2 , respectively.

Specifically, we will derive explicit upper and lower bounds on $R^*(n, \epsilon)$ in terms of the first three moments of the information random variable $\imath_X(X)$. Although particularizing Theorem 6 it is possible, in principle, to compute $R^*(n, \epsilon)$ exactly, it is more desirable to derive approximations that are both easier to compute and offer more intuition into the behavior of the fundamental limit $R^*(n, \epsilon)$.

Theorems 16 and 17 imply that, for all $\epsilon \in (0, 1/2)$, the best achievable rate $R^*(n, \epsilon)$ satisfies,

$$\frac{c}{n} \leq R^*(n, \epsilon) - \left[H + \frac{\sigma}{\sqrt{n}} Q^{-1}(\epsilon) - \frac{\log_2 n}{2n} \right] \leq \frac{c'}{n}. \quad (121)$$

The upper bound is valid for all n , the lower bound is valid for $n \geq n_0$ as in (157), and explicit values are derived for the constants c, c' . In view of Theorem 1, essentially the same results as in (121) hold for prefix codes,

$$\frac{c}{n} \leq R_p(n, \epsilon) - \left[H + \frac{\sigma}{\sqrt{n}} Q^{-1}(\epsilon) - \frac{\log_2 n}{2n} \right] \leq \frac{c' + 1}{n}. \quad (122)$$

The bounds in equations (121) and (122) justify the Gaussian approximation (37) stated in Section I.

Before establishing the precise non-asymptotic relations leading to (121) and (122), we illustrate their utility via an example. To facilitate this, note that Theorem 2 immediately yields the following simple bound:

Theorem 14: For all $n \geq 1$, $\epsilon > 0$,

$$R^*(n, \epsilon) \leq R^u(n, \epsilon), \quad (123)$$

where $R^u(n, \epsilon)$ is the quantile function of the information spectrum, i.e., the lowest R such that:

$$\mathbb{P} \left[\frac{1}{n} \sum_{i=1}^n \imath_X(X_i) \geq R \right] \leq \epsilon. \quad (124)$$

In Figure 2, we exhibit the behavior of the fundamental compression limit $R^*(n, \epsilon)$ in the case of coin flips with bias 0.11 (for which $H \approx 0.5$ bits). In particular, we compare $R^*(n, \epsilon)$ and $R^u(n, \epsilon)$ for $\epsilon = 0.1$. The non-monotonic nature of both $R^*(n, \epsilon)$ and $R^u(n, \epsilon)$ with n is not surprising: although the larger the value of n the less we are at the mercy of the source randomness, we also need to compress more information. Figure 2 also illustrates that $R^*(n, \epsilon)$ is tracked rather closely by the Gaussian approximation,

$$\tilde{R}^*(n, \epsilon) = H + Q^{-1}(\epsilon) \frac{\sigma}{\sqrt{n}} - \frac{1}{2n} \log_2 n, \quad (125)$$

suggested by (121).

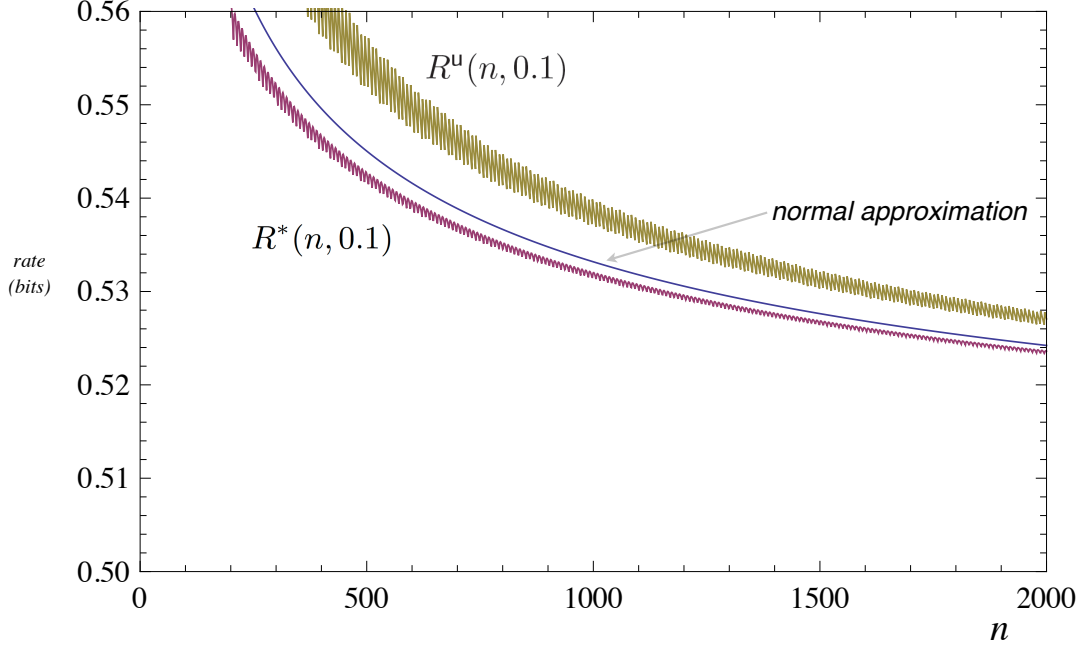


Fig. 2: The optimum rate $R^*(n, 0.1)$, the Gaussian approximation $\tilde{R}^*(n, 0.1)$ in (125), and the upper bound $R^u(n, 0.1)$, for a Bernoulli-0.11 source and blocklengths $200 \leq n \leq 2000$.

Figure 3 focuses the comparison between $R^*(n, 0.1)$ and $\tilde{R}^*(n, 0.1)$ on the short blocklength range up to 200 not shown in Figure 2. For $n > 60$, the discrepancy between the two never exceeds 4%.

The remainder of the section is devoted to justifying the use of (125) as an accurate approximation to $R^*(n, \epsilon)$. To that end, in Theorems 17 and 16 we establish the bounds given in (121). Their derivation requires that we overcome two technical hurdles:

- 1) The distribution function of the optimal encoding length is not the same as the distribution of $\frac{1}{n} \sum_{i=1}^n \iota_X(X_i)$;
- 2) The distribution of $\frac{1}{n} \sum_{i=1}^n \iota_X(X_i)$ is only approximately Gaussian.

To cope with the second hurdle we will appeal to the classical Berry-Esséen bound [15], [21]:

Theorem 15: Let $\{Z_i\}$ be independent and identically distributed random variables with zero mean and unit variance, and let \bar{Z} be standard normal. Then, for all $n \geq 1$ and any a :

$$\left| \mathbb{P} \left[\frac{1}{\sqrt{n}} \sum_{i=1}^n Z_i \leq a \right] - \mathbb{P} [\bar{Z} \leq a] \right| \leq \frac{\mathbb{E}[|Z_1 - \mathbb{E}[Z_1]|^3]}{2\sqrt{n}}. \quad (126)$$

Invoking the Berry-Esséen bound, Strassen [26] claimed the following approximation for $n > \frac{19600}{\delta^{16}}$,

$$\left| R^*(n, \epsilon) - \tilde{R}^*(n, \epsilon) \right| \leq \frac{140}{\delta^8}, \quad (127)$$

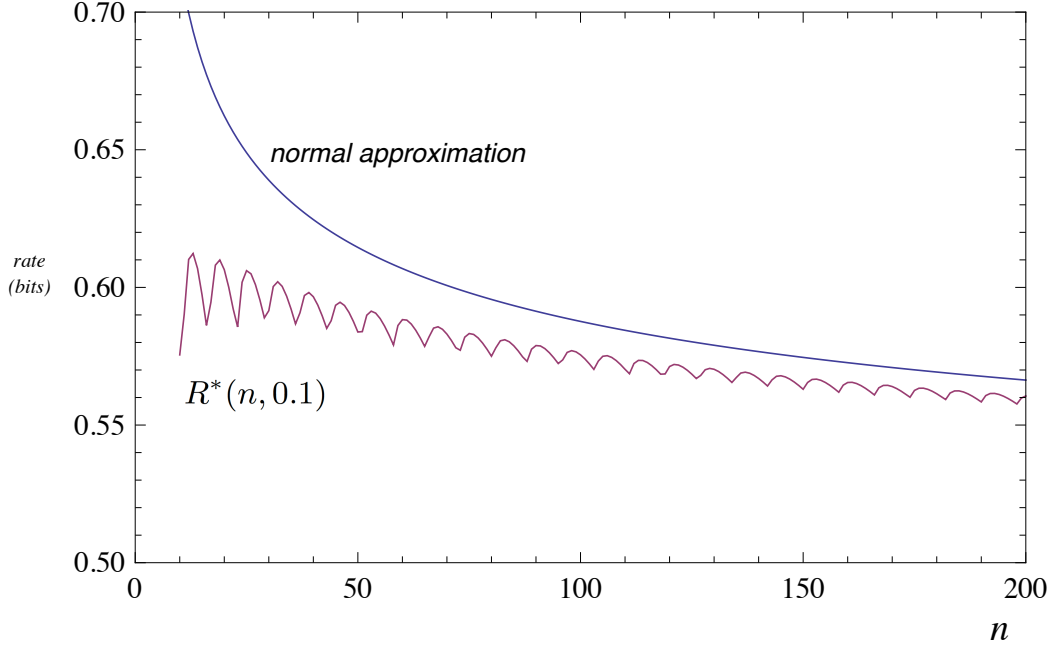


Fig. 3: The optimum rate $R^*(n, 0.1)$ and the Gaussian approximation $\tilde{R}^*(n, 0.1)$ in (125), for a Bernoulli-0.11 source and blocklengths $10 \leq n \leq 200$.

where

$$\delta \leq \min \left\{ \sigma, \epsilon, 1 - \epsilon, \mu_3^{-1/3} \right\} \quad (128)$$

$$\mu_3 = \mathbb{E}[|\iota_X(X) - H|^3]. \quad (129)$$

Unfortunately, we were not able to verify how [26] justifies the application of (126) to bound integrals with respect to the corresponding cumulative distribution functions (cf. equations (2.17), (3.18) and the displayed equation between (3.15) and (3.16) in [26]).

The following achievability result holds for all blocklengths.

Theorem 16: For all $0 < \epsilon \leq \frac{1}{2}$ and all $n \geq 1$,

$$\begin{aligned} R^*(n, \epsilon) \leq & H + \frac{\sigma}{\sqrt{n}} Q^{-1}(\epsilon) - \frac{\log_2 n}{2n} \\ & + \frac{1}{n} \log_2 \left(\frac{\log_2 e}{\sqrt{2\pi\sigma^2}} + \frac{\mu_3}{\sigma^3} \right) \\ & + \frac{1}{n \sigma^2 \phi(\Phi^{-1}(\Phi(Q^{-1}(\epsilon)) + \frac{\mu_3}{\sigma^3 \sqrt{n}}))}, \end{aligned} \quad (130)$$

as long as the varentropy σ^2 is strictly positive, where $\Phi = 1 - Q$ and $\phi = \Phi'$ are the standard Gaussian distribution function and density, respectively.

Proof: The proof follows Strassen's construction, but the essential approximation steps are different. The positive constant β_n is uniquely defined by:

$$\mathbb{P}[\iota_{X^n}(X^n) \leq \log_2 \beta_n] \geq 1 - \epsilon, \quad (131)$$

$$\mathbb{P}[\iota_{X^n}(X^n) < \log_2 \beta_n] < 1 - \epsilon. \quad (132)$$

Since the information spectrum (i.e., the distribution function of the information random variable $\iota_{X^n}(X^n)$) is piecewise constant, $\log_2 \beta_n$ is the location of the jump where the information spectrum reaches (or exceeds for the first time) the value $1 - \epsilon$. Furthermore, defining the normalized constant,

$$\lambda_n = \frac{\log_2 \beta_n - nH}{\sqrt{n}\sigma}, \quad (133)$$

the probability in the left side of (131) is,

$$\mathbb{P}\left[\frac{\iota_{X^n}(X^n) - nH}{\sqrt{n}\sigma} \leq \lambda_n\right] \leq \Phi(\lambda_n) + \frac{\mu_3}{2\sigma^3\sqrt{n}}, \quad (134)$$

where we have applied Theorem 15. Analogously, we obtain,

$$\mathbb{P}\left[\frac{\iota_{X^n}(X^n) - nH}{\sqrt{n}\sigma} < \lambda_n\right] \geq \Phi(\lambda_n) - \frac{\mu_3}{2\sigma^3\sqrt{n}}. \quad (135)$$

Since $1 - \epsilon$ is sandwiched between the right sides of (134) and (135), as $n \rightarrow \infty$ we must have that, $\lambda_n \rightarrow \lambda$, where,

$$\lambda = \Phi^{-1}(1 - \epsilon) = Q^{-1}(\epsilon). \quad (136)$$

By a simple first-order Taylor bound,

$$\lambda_n \leq \Phi^{-1}\left(\Phi(\lambda) + \frac{\mu_3}{2\sigma^3\sqrt{n}}\right) \quad (137)$$

$$= \lambda + \frac{\mu_3}{2\sigma^3\sqrt{n}}(\Phi^{-1})'(\xi_n) \quad (138)$$

$$= \lambda + \frac{\mu_3}{2\sigma^3\sqrt{n}} \frac{1}{\phi(\Phi^{-1}(\xi_n))}, \quad (139)$$

for some $\xi_n \in [\Phi(\lambda), \Phi(\lambda) + \frac{\mu_3}{2\sigma^3\sqrt{n}}]$. Since $\epsilon \leq 1/2$, we have $\lambda \geq 0$ and $\Phi(\lambda) \geq 1/2$, so that $\xi_n \geq 1/2$. And since $\Phi^{-1}(t)$ is strictly increasing for all t , while ϕ is strictly decreasing for $t \geq 0$, from (139) we obtain,

$$\lambda_n \leq \lambda + \frac{\mu_3}{2\sigma^3\sqrt{n}} \frac{1}{\phi(\Phi^{-1}(\Phi(\lambda) + \frac{\mu_3}{2\sigma^3\sqrt{n}}))}. \quad (140)$$

The event E_n in the left side of (131) contains all the “high probability strings,” and itself it has probability $\geq 1 - \epsilon$. Its cardinality is $M_X^+(\beta_n)$, defined in (75) (with $X \leftarrow X^n$). Therefore, denoting,

$$p(t) = 2^{-t} \mathbb{I}\{t \geq 0\} \quad (141)$$

$$Y_i = \frac{1}{\sigma} (\iota_X(X_i) - H), \quad (142)$$

we obtain,

$$R^*(n, \epsilon) \leq \frac{1}{n} \log_2 M_X^+(\beta_n) \quad (143)$$

$$= \frac{1}{n} \log_2 \mathbb{E} [\exp(\iota_{X^n}(X^n)) \mathbb{1}_{\{\iota_{X^n}(X^n) \leq \log_2 \beta_n\}}] \quad (144)$$

$$= H + \lambda_n \frac{\sigma}{\sqrt{n}} + \frac{1}{n} \log_2 \alpha_n, \quad (145)$$

with,

$$\alpha_n = \mathbb{E} [p(\log_2 \beta_n - \iota_{X^n}(X^n))] \quad (146)$$

$$= \mathbb{E} \left[p \left(\sqrt{n} \sigma \left(\lambda_n - \frac{1}{\sqrt{n}} \sum_{i=1}^n Y_i \right) \right) \right], \quad (147)$$

and where the $\{Y_i\}$ are independent, identically distributed, with zero mean and unit variance. Let $\bar{\alpha}_n$ be defined as (147) except that Y_i are replaced by \bar{Y}_i which are standard normal. Then, straightforward algebra yields,

$$\bar{\alpha}_n = \mathbb{E} \left[2^{-\sqrt{n}\sigma(\lambda_n - \bar{Y}_1)} \mathbb{I}\{\bar{Y}_1 \leq \lambda_n\} \right] \quad (148)$$

$$= \int_0^\infty 2^{-x} \frac{e^{-\frac{(x + \lambda_n \sigma \sqrt{n})^2}{2\sigma^2 n}}}{\sqrt{2\pi\sigma^2 n}} dx \quad (149)$$

$$\leq \frac{\log_2 e}{\sqrt{2\pi\sigma^2 n}}. \quad (150)$$

To deal with the fact that the random variables in (147) are not normal, we apply the Lebesgue-Stieltjes integration by parts formula to (147). Denoting the distribution of the normalized sum in (147) by $F_n(t)$, α_n becomes,

$$\alpha_n = \int_{-\infty}^{\lambda_n} 2^{-(\sqrt{n}\sigma(\lambda_n - t))} dF_n(t) \quad (151)$$

$$= F_n(\lambda_n) - \int_{-\infty}^{\lambda_n} F_n(t) \sqrt{n}\sigma 2^{-(\sqrt{n}\sigma(\lambda_n - t))} dt \log_e 2 \quad (152)$$

$$= \bar{\alpha}_n + F_n(\lambda_n) - \Phi(\lambda_n) - \int_{-\infty}^{\lambda_n} (F_n(t) - \Phi(t)) \sqrt{n}\sigma 2^{-(\sqrt{n}\sigma(\lambda_n - t))} dt \log_e 2 \quad (153)$$

$$\leq \bar{\alpha}_n + \frac{\mu_3}{2\sigma^3\sqrt{n}} + \frac{\mu_3}{2\sigma^2} \int_{-\infty}^{\lambda_n} 2^{-(\sqrt{n}\sigma(\lambda_n - t))} dt \log_e 2 \quad (154)$$

$$= \bar{\alpha}_n + \frac{\mu_3}{\sigma^3\sqrt{n}} \quad (155)$$

$$\leq \frac{1}{\sqrt{n}} \left(\frac{\log_2 e}{\sqrt{2\pi\sigma^2}} + \frac{\mu_3}{\sigma^3} \right), \quad (156)$$

where (154) follows from Theorem 15. The desired result now follows from (145) after assembling the bounds on λ_n and α_n in (140) and (156), respectively. \blacksquare

Next we give a complementary converse result.

Theorem 17: For all $0 < \epsilon < \frac{1}{2}$ and all n such that,

$$n > n_0 = \frac{1}{4} \left(1 + \frac{\mu_3}{2\sigma^3}\right)^2 \frac{1}{(\phi(Q^{-1}(\epsilon))Q^{-1}(\epsilon))^2}, \quad (157)$$

the following lower bound holds,

$$R^*(n, \epsilon) \geq H + \frac{\sigma}{\sqrt{n}} Q^{-1}(\epsilon) - \frac{\log_2 n}{2n} - \frac{\frac{\mu_3}{2} + \sigma^3}{n\sigma^2\phi(Q^{-1}(\epsilon))}, \quad (158)$$

as long as the varentropy σ^2 is strictly positive.

Proof: Let,

$$\eta = \frac{\frac{\mu_3}{2\sigma^2} + \sigma}{\phi(Q^{-1}(\epsilon))}, \quad (159)$$

and consider,

$$\begin{aligned} & \mathbb{P} \left[\sum_{i=1}^n \iota_X(X_i) \geq Hn + \sigma\sqrt{n}Q^{-1}(\epsilon) - \eta \right] \\ &= \mathbb{P} \left[\sum_{i=1}^n \frac{\iota_X(X_i) - H}{\sigma\sqrt{n}} \geq Q^{-1}(\epsilon) - \frac{\eta}{\sigma\sqrt{n}} \right] \end{aligned} \quad (160)$$

$$\geq Q \left(Q^{-1}(\epsilon) - \frac{\eta}{\sigma\sqrt{n}} \right) - \frac{\mu_3}{2\sigma^3\sqrt{n}} \quad (161)$$

$$\geq \epsilon + \frac{\eta}{\sigma\sqrt{n}}\phi(Q^{-1}(\epsilon)) - \frac{\mu_3}{2\sigma^3\sqrt{n}} \quad (162)$$

$$= \epsilon + \frac{1}{\sqrt{n}}, \quad (163)$$

where (161) follows from Theorem 15, and (162) follows from,

$$Q(a - \Delta) \geq Q(a) + \Delta\phi(Q(a)), \quad (164)$$

which holds at least as long as,

$$a > \frac{\Delta}{2} > 0. \quad (165)$$

Letting $a = Q^{-1}(\epsilon)$ and $\Delta = \frac{\eta}{\sigma\sqrt{n}}$, (165) is equivalent to (157).

We proceed to invoke Theorem 3 with $X \leftarrow X^n$, k equal to n times the right side of (158), and $\tau = \frac{1}{2} \log_2 n$. In view of the definition of $R^*(n, \epsilon)$ and (160)-(163), the desired result follows. \blacksquare

VI. GAUSSIAN APPROXIMATION FOR MARKOV SOURCES

Let $\{X_n\}$ be an irreducible, aperiodic, k th order Markov chain on the finite alphabet \mathcal{A} , with transition probabilities,

$$P_{X'|X^k}(x_{k+1} | x^k), \quad x^{k+1} \in \mathcal{A}^{k+1}, \quad (166)$$

and entropy rate H . Note that we do not assume that the source is stationary. In Theorem 12 of Section IV we established that the varentropy rate defined in general in equation (108), for stationary ergodic chains exists as the limit,

$$\sigma^2 = \lim_{n \rightarrow \infty} \frac{1}{n} \text{Var}(\iota_{X^n}(X^n)). \quad (167)$$

An examination of the proof shows that, by an application of the general central limit theorem for (uniformly ergodic) Markov chains [5], [19], the assumption of stationarity is not necessary, and (167) holds for all irreducible aperiodic chains.

Theorem 18: Suppose $\{X_n\}$ is an irreducible and aperiodic k th order Markov source, and let $\epsilon \in (0, 1/2)$. Then, there is a positive constant C such that, for all n large enough,

$$nR^*(n, \epsilon) \leq nH + \sigma\sqrt{n}Q^{-1}(\epsilon) + C, \quad (168)$$

where the varentropy rate σ^2 is given by (167) and it is assumed to be strictly positive.

Theorem 19: Under the same assumptions as in Theorem 18, for all n large enough,

$$nR^*(n, \epsilon) \geq nH + \sigma\sqrt{n}Q^{-1}(\epsilon) - \frac{1}{2} \log_2 n - C, \quad (169)$$

where $C > 0$ is a finite constant, possibly different from than in Theorem 18.

Remarks.

- 1) By definition, the lower bound in Theorem 19 also applies to $R_p(n, \epsilon)$, while in view Theorem 1, the upper bound in Theorem 18 also applies to $R_p(n, \epsilon)$ provided C is replaced by $C + 1$.
- 2) Note that, unlike the direct and converse coding theorems for memoryless sources (Theorems 16 and 17, respectively) the results of Theorems 18 and 19 are asymptotic in that we do not give explicit bounds for the constant terms. This is because the main probabilistic tool we use in the proofs (the Berry-Esséen bound in Theorem 15) does not have an equally precise counterpart for Markov chains. Specifically, in the proof of Theorem 20 below we appeal to a Berry-Esséen bound established by Nagaev in [20], which does not give an explicit value for the multiplicative constant A in (174). More explicit bounds do exist, but they require additional conditions on the Markov chain; see, e.g., Mann's thesis [16], and the references therein.
- 3) If we restrict our attention to the (much more narrow) class of *reversible* chains, then it is indeed possible to apply the Berry-Esséen bound of Mann [16] to obtain explicit values for the constants in Theorems 18 and 19; but the resulting values are pretty loose, drastically limiting the engineering usefulness of the resulting bounds. For example, in Mann's version of the Berry-Esséen bound, the corresponding right side of the inequality as in Theorem 15 is multiplied by a factor of 13000. Therefore, we have opted for the less explicit but much more general statements given above.

- 4) Similar comments to those in the last two remarks apply to the observation that Theorem 18 is a weaker bound than that established in Theorem 16 for memoryless sources, by a $(1/2) \log_2 n$ term. Instead of restricting our result to the much more narrow class of reversible chains, or extending the involved proof of Theorem 16 to the case of a Markovian source, we chose to illustrate how this slightly weaker bound can be established in full generality, with a much shorter and simpler proof.
- 5) The proof of Theorem 18 shows that the constant in its statement can be chosen as

$$C = \frac{2A\sigma}{\phi(Q^{-1}(\epsilon))} \quad (170)$$

for all

$$n \geq \frac{8A^2}{\pi e (\phi(Q^{-1}(\epsilon)))^4}, \quad (171)$$

where A is the constant appearing in Theorem 20, below. Similarly, from the proof of Theorem 19 we see that the constant in its statement can be chosen as

$$C = \frac{\sigma(A+1)}{\phi(Q^{-1}(\epsilon))} + 1, \quad (172)$$

for all,

$$n \geq \left(\frac{A+1}{Q^{-1}(\epsilon)\phi(Q^{-1}(\epsilon))} \right)^2. \quad (173)$$

Note that, in both cases, the values of the constants can easily be improved, but they still depend on the implicit constant A of Theorem 20.

As mentioned above, we will need a Berry-Esséen-type bound on the scaled information random variables,

$$\frac{\iota_{X^n}(X^n) - nH}{\sqrt{n}\sigma}.$$

Beyond the Shannon-McMillan-Breiman theorem, several more refined asymptotic results have been established for this sequence; see, in particular, [8], [22], [26], [33] and the discussions in [12] and in Section IV. Unlike these asymptotic results, we will use the following non-asymptotic bound.

Theorem 20: For an ergodic, k th order Markov source $\{X_n\}$ with entropy rate H and positive varentropy rate σ^2 , there exists a finite constant $A > 0$ such that, for all $n \geq 1$,

$$\sup_{z \in \mathbb{R}} \left| \mathbb{P} \left[\iota_{X^n}(X^n) - nH > z \sigma \sqrt{n} \right] - Q(z) \right| \leq \frac{A}{\sqrt{n}}. \quad (174)$$

Proof: For integers $i \leq j$, we adopt the notation x_i^j and X_i^j for blocks of strings $(x_i, x_{i+1}, \dots, x_j)$ and random variables $(X_i, X_{i+1}, \dots, X_j)$, respectively. For all $x^{n+k} \in \mathcal{A}^{n+k}$

such that $P_{X^k}(x^k) > 0$ and $P_{X'|X_{j-k}^{j-1}}(x_j | x_{j-k}^{j-1}) > 0$, for $j = k+1, k+2, \dots, n+k$, we have,

$$\iota_{X^n}(x^n) = \log_2 \frac{1}{P_{X^k}(x^k) \prod_{j=k+1}^n P_{X'|X_{j-k}^{j-1}}(x_j | x_{j-k}^{j-1})} \quad (175)$$

$$\begin{aligned} &= \sum_{j=k+1}^{k+n} \log_2 \frac{1}{P_{X'|X_{j-k}^{j-1}}(x_j | x_{j-k}^{j-1})} \\ &- \log_2 \frac{P_{X^k}(x^k)}{\prod_{j=n+1}^{n+k} P_{X'|X_{j-k}^{j-1}}(x_j | x_{j-k}^{j-1})} \end{aligned} \quad (176)$$

$$= \sum_{j=1}^n f(x^{j+k}) + \Delta_n, \quad (177)$$

where the function $f: A' \rightarrow \mathbb{R}$ is defined by,

$$f(x^{k+1}) = \iota_{X'|X^k}(x_{k+1} | x^k) = \log_2 \frac{1}{P_{X'|X^k}(x_{k+1} | x^k)}, \quad (178)$$

and,

$$\Delta_n = -\log_2 \frac{P_{X^k}(x^k)}{\prod_{j=n+1}^{n+k} P_{X'|X_{j-k}^{j-1}}(x_j | x_{j-k}^{j-1})}. \quad (179)$$

Denote

$$|\Delta_n| \leq \delta = \max \left| \log_2 \left[\frac{P_{X^k}(x^k)}{\prod_{j=n+1}^{n+k} P_{X'|X^k}(x_j | x_{j-k}^{j-1})} \right] \right| < \infty, \quad (180)$$

where the maximum is over the positive probability strings for which we have established (177).

Let $\{Y_n\}$ denote the first-order Markov source defined by taking overlapping $(k+1)$ -blocks in the original chain,

$$Y_n = (X_n, X_{n+1}, \dots, X_{n+k}). \quad (181)$$

Since $\{X_n\}$ is irreducible and aperiodic, so is $\{Y_n\}$ on the state space,

$$\mathcal{A}' = \{x^{k+1} \in \mathcal{A}^{k+1} : P_{X'|X^k}(x_{k+1} | x^k) > 0\}. \quad (182)$$

Now, since the chain $\{Y_n\}$ is irreducible and aperiodic on a finite state space, condition (0.2) of [20] is satisfied, and since the function f is bounded, Theorem 1 of [20] implies that there exists a finite constant A_1 such that, for all n ,

$$\sup_{z \in \mathbb{R}} \left| \mathbb{P} \left[\frac{\sum_{j=1}^n f(Y_j) - nH}{\sigma \sqrt{n}} > z \right] - Q(z) \right| \leq \frac{A_1}{\sqrt{n}}, \quad (183)$$

where the entropy rate is $H = \mathbb{E}[f(\tilde{Y}_1)]$ and,

$$\Sigma^2 = \lim_{n \rightarrow \infty} \frac{1}{n} \mathbb{E} \left[\left(\sum_{j=1}^n (f(\tilde{Y}_j) - H) \right)^2 \right], \quad (184)$$

where $\{\tilde{Y}_n\}$ is a stationary version of $\{Y_n\}$, that is, it has the same transition probabilities but its initial distribution is its unique invariant distribution,

$$\mathbb{P}[\tilde{Y}_1 = x^{k+1}] = \pi(x^k)P_{X'|X^k}(x_{k+1} | x^k), \quad (185)$$

where π is the unique invariant distribution of the original chain $\{X_n\}$. Since the function f is bounded and the distribution of the chain $\{Y_n\}$ converges to stationarity exponentially fast, it is easy to see that (184) coincides with the source varentropy rate.

Let $F_n(z)$, $G_n(z)$ denote the complementary cumulative distribution functions,

$$F_n(z) = \mathbb{P}[\iota_{X^n}(X^n) - nH > z\sqrt{n}\sigma], \quad (186)$$

$$G_n(z) = \mathbb{P}\left[\sum_{j=1}^n f(Y_j) - nH > z\sqrt{n}\sigma\right]. \quad (187)$$

Since $F_n(z)$ and $G_n(z)$ are non-increasing, (177) and (180) imply that

$$F_n(z) \geq G_n(z + \delta/\sqrt{n}) \quad (188)$$

$$\geq Q(z + \delta/\sqrt{n}) - \frac{A_1}{\sqrt{n}}, \quad (189)$$

$$\geq Q(z) - \frac{A}{\sqrt{n}}, \quad (190)$$

uniformly in z , where (189) follows from (184), and (190) holds with $A = A_1 + \delta/\sqrt{2\pi}$ since $Q'(z) = -\phi(z)$ is bounded by $-1/\sqrt{2\pi}$. A similar argument shows that,

$$F_n(z) \leq G_n(z - \delta/\sqrt{n}) \quad (191)$$

$$\leq Q(z - \delta/\sqrt{n}) + \frac{A_1}{\sqrt{n}} \quad (192)$$

$$\leq Q(z) + \frac{A}{\sqrt{n}}. \quad (193)$$

Since both (190) and (193) hold uniformly in $z \in \mathbb{R}$, together they form the statement of the theorem. \blacksquare

Proof of Theorem 18: Starting from Theorem 2 with X^n in place of X and with,

$$K_n = nH + \sigma\sqrt{n}Q^{-1}(\epsilon) + C, \quad (194)$$

where C will be chosen below, Theorem 2 states that,

$$\mathbb{P}[\ell(f_n^*(X^n)) \geq K_n] \leq \mathbb{P}[\iota_{X^n}(X^n) \geq K_n] \quad (195)$$

$$= \mathbb{P}\left[\frac{1}{\sigma\sqrt{n}}(\iota_{X^n}(X^n) - nH) \geq Q^{-1}(\epsilon) + \frac{C}{\sigma\sqrt{n}}\right] \quad (196)$$

$$\leq Q\left(Q^{-1}(\epsilon) + \frac{C}{\sigma\sqrt{n}}\right) + \frac{A}{\sqrt{n}}, \quad (197)$$

where (197) follows from Theorem 20. Since,

$$Q'(x) = -\phi(x) \quad (198)$$

$$0 \leq Q''(x) = x\phi(x) \leq \frac{1}{\sqrt{2\pi}e}, \quad x \geq 0, \quad (199)$$

a second-order Taylor expansion of the first term in the right side of (197) gives,

$$\mathbb{P}[\ell(\mathbf{f}_n^*(X^n)) \geq K_n] \leq \epsilon - \frac{C}{\sigma\sqrt{n}}\phi(Q^{-1}(\epsilon)) + \frac{1}{2\sqrt{2\pi e}}\left(\frac{C}{\sigma\sqrt{n}}\right)^2 + \frac{A}{\sqrt{n}} \quad (200)$$

$$\leq \epsilon - \frac{1}{\sigma\sqrt{n}} \left\{ C \left[\phi(Q^{-1}(\epsilon)) - \frac{1}{2\sqrt{2\pi e}} \left(\frac{C}{\sigma\sqrt{n}} \right) \right] - A\sigma \right\}, \quad (201)$$

and choosing C as in (170) for n satisfying (171) the right side of (201) is bounded above by ϵ . Therefore, $\mathbb{P}[\ell(\mathbf{f}_n^*(X^n)) > K_n] \leq \epsilon$, which, by definition implies that $nR^*(n, \epsilon) \leq K_n$, as claimed. \blacksquare

Proof of Theorem 19: Applying Theorem 3 with X^n in place of X and with $\delta > 0$ and $K_n \geq 1$ arbitrary, we obtain,

$$\mathbb{P}[\ell(\mathbf{f}_n^*(X^n)) \geq K_n] \geq \mathbb{P}[\iota_{X^n}(X^n) \geq K_n + \delta] - 2^{-\delta} \quad (202)$$

$$= \mathbb{P}\left[\frac{1}{\sigma\sqrt{n}}(\iota_{X^n}(X^n) - nH) \geq \frac{K_n - nH + \delta}{\sigma\sqrt{n}}\right] - 2^{-\delta} \quad (203)$$

$$\geq Q\left(\frac{K_n - nH + \delta}{\sigma\sqrt{n}}\right) - \frac{A}{\sqrt{n}} - 2^{-\delta}, \quad (204)$$

where (204) now follows from Theorem 20. Letting $\delta = \delta_n = \frac{1}{2} \log_2 n$ and,

$$K_n = nH + \sigma\sqrt{n}Q^{-1}(\epsilon) - \delta - \frac{\sigma(A+1)}{\phi(Q^{-1}(\epsilon))}, \quad (205)$$

yields,

$$\mathbb{P}[\ell(\mathbf{f}_n^*(X^n)) \geq K_n] \geq Q\left(Q^{-1}(\epsilon) - \frac{(A+1)}{\phi(Q^{-1}(\epsilon))\sqrt{n}}\right) - \frac{1}{\sqrt{n}}. \quad (206)$$

Note that, since $\epsilon \in (0, 1/2)$, we have $Q^{-1}(\epsilon) > 0$. And since $Q'(x) = -\phi(x)$, a simple two-term Taylor expansion of Q above gives,

$$\mathbb{P}[\ell(\mathbf{f}_n^*(X^n)) \geq K_n] \geq \epsilon + \frac{A}{\sqrt{n}} > \epsilon, \quad (207)$$

for all,

$$n \geq \left(\frac{A+1}{Q^{-1}(\epsilon)\phi(Q^{-1}(\epsilon))} \right)^2,$$

hence $nR^*(n, \epsilon) > K_n - 1$, as claimed. \blacksquare

VII. SOURCE DISPERSION AND VARENTROPY

Traditionally, refined analyses in lossless data compression have focused attention on the *redundancy*, defined as the difference between the minimum average compression rate and the entropy rate. As we mentioned in Section I-D, if the source statistics are known, then the per-symbol redundancy is positive and behaves as $O(\frac{1}{n})$ when the prefix condition is enforced, while it is $-\frac{1}{2n} \log_2 n + O(\frac{1}{n})$, without the prefix condition. But since, as we saw in Sections V and VI, the standard deviation of the best achievable compression rate is $O(\frac{1}{\sqrt{n}})$, the rate will be dominated by these fluctuations. Therefore, as noted in [11], it of primary importance to analyze the variance of the optimal codelengths. To that end, we introduce the following operational definition:

Definition 2: The dispersion D (measured in bits²) of a source $\{P_{X^n}\}_{n=1}^\infty$ is,

$$D = \limsup_{n \rightarrow \infty} \frac{1}{n} \text{Var}(\ell(f_n^*(X^n))), \quad (208)$$

where $\ell(f_n^*(\cdot))$ is the length of the optimum fixed-to-variable lossless code (cf. Section I-A).

As we show in Theorem 22 below, for a broad class of sources, the dispersion D is equal to the source varentropy rate σ^2 defined in (108). Moreover, in view of the Gaussian approximation bounds for $R^*(n, \epsilon)$ in Sections V and VI – and more generally, as long as a similar two-term Gaussian approximation in terms of the entropy rate and varentropy rate can be established up to $o(1/\sqrt{n})$ accuracy – we can conclude the following: by the definition of $n^*(R, \epsilon)$ in Section I-A, the source blocklength n required for the compression rate to exceed $(1 + \eta)H$ with probability no greater than $\epsilon > 0$ is approximated by,

$$n^*((1 + \eta)H, \epsilon) \approx \frac{\sigma^2}{H^2} \left(\frac{Q^{-1}(\epsilon)}{1 + \eta} \right)^2 \quad (209)$$

$$= \frac{D}{H^2} \left(\frac{Q^{-1}(\epsilon)}{1 + \eta} \right)^2, \quad (210)$$

i.e., by the product of a factor that depends only on the source (through H and D or σ^2), and a factor that depends only on the design requirements ϵ and η . Note that this is in close parallel with the notion of channel dispersion introduced in [23].

Example 4: Coin flips with bias p have varentropy,

$$\sigma^2 = p(1 - p) \log^2 \frac{1 - p}{p}, \quad (211)$$

so the key parameter in (210) which characterizes the time horizon required for the source to become “typical” is,

$$\frac{D}{H^2} = \frac{p - p^2}{\left(p + \frac{1}{\frac{\log p}{\log(1-p)} - 1} \right)^2}. \quad (212)$$

Example 5: For a memoryless source whose marginal is the geometric distribution,

$$P_X(k) = q(1 - q)^k, \quad (213)$$

the ratio of varentropy to squared entropy is,

$$\frac{\sigma^2}{H^2} = \frac{\sigma^2}{H^2} = (1 - q) \left(\frac{\log_2(1 - q)}{h(q)} \right)^2, \quad (214)$$

where h denotes the binary entropy function.

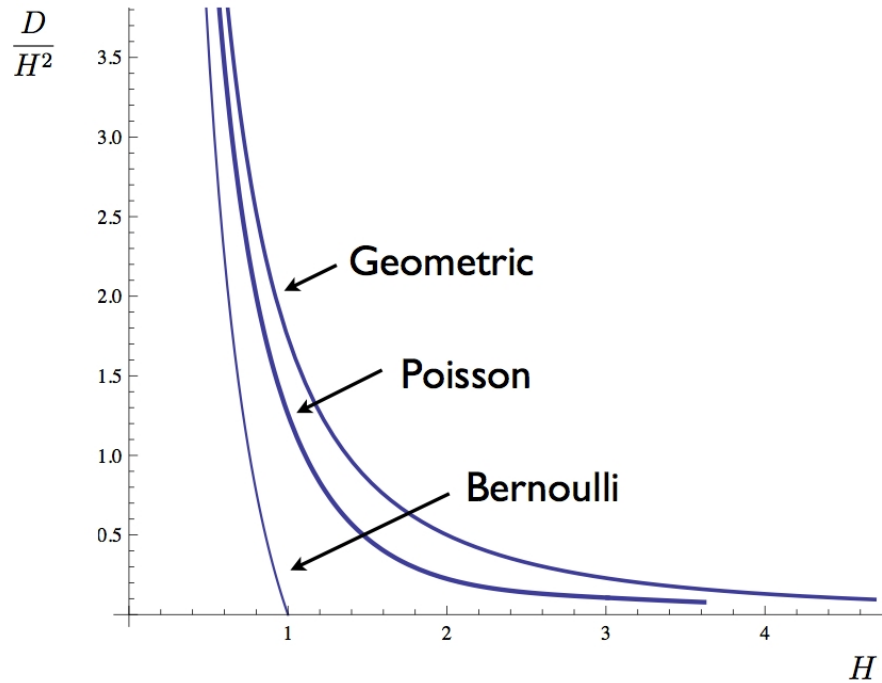


Fig. 4: Normalized dispersion as a function of entropy for memoryless sources

Figure 4 compares the normalized dispersion to the entropy for the Bernoulli, geometric and Poisson distributions. We see that as the source becomes more compressible (lower entropy per letter), the longer the horizon over which we need to compress in order to squeeze most of the redundancy out of the source.

Definition 3: A source $\{X_n\}$ taking values on the finite alphabet \mathcal{A} is a *linear information growth* source if any nonzero-probability string has probability bounded below by an exponential, that is, if there is a finite constant A and an integer $N_0 \geq 1$ such that, for all $n \geq N_0$, every nonzero-probability string $x^n \in \mathcal{A}^n$ satisfies

$$\mathbb{P}_{X^n}(x^n) \leq An. \quad (215)$$

Any memoryless source belongs to the class of linear information growth. Also note that, every irreducible and aperiodic Markov chain is a linear information growth source: Writing q for the smallest nonzero element of the transition matrix, and π for the smallest nonzero probability for X_1 , we easily see that (215) is satisfied with $N_0 = 1$, $A = \log_2(1/q\pi)$. The

class of linear information growth sources is related, at least at the level of intuition, to the class of finite-energy processes considered by Shields [25] and to processes satisfying the Doeblin-like condition of Kontoyiannis and Suhov [14].

We proceed to show an interesting regularity result for linear information growth sources:

Lemma 1: Suppose $\{X_n\}$ is a (not necessarily stationary or ergodic) linear information growth source. Then:

$$\lim_{n \rightarrow \infty} \frac{1}{n} \mathbb{E} \left[\left(\ell(f_n^*(X^n)) - \imath_{X^n}(X^n) \right)^2 \right] = 0. \quad (216)$$

Proof: For brevity, denote $\ell_n = \ell(f_n^*(X^n))$ and $\imath_n = \imath_{X^n}(X^n)$, respectively. Select an arbitrary τ_n . The expectation of interest is

$$\mathbb{E}[(\ell_n - \imath_n)^2] = \mathbb{E}[(\ell_n - \imath_n)^2 \mathbb{I}\{\ell_n \geq \imath_n - \tau_n\}] + \mathbb{E}[(\ell_n - \imath_n)^2 \mathbb{I}\{\ell_n < \imath_n - \tau_n\}]. \quad (217)$$

Since $\ell_n \leq \imath_n$, on the event $\{\ell_n \geq \imath_n - \tau_n\}$, we have $(\ell_n - \imath_n)^2 \leq \tau_n^2$. Also, by the linear information growth assumption we have the bound $0 \leq \imath_n - \ell_n \leq \imath_n \leq Cn$ for a finite constant C and all n large enough. Combining these two observations with Theorem 4, we obtain that,

$$\mathbb{E}[(\ell_n - \imath_n)^2] \leq \tau_n^2 + C^2 n^2 \mathbb{P}\{\ell_n < \imath_n - \tau_n\} \quad (218)$$

$$\leq \tau_n^2 + C^2 n^2 2^{-\tau_n} (n \log_2 |\mathcal{A}| + 1) \quad (219)$$

$$\leq \tau_n^2 + C' n^3 2^{-\tau_n}, \quad (220)$$

for some $C' < \infty$ and all n large enough. Taking $\tau_n = 3 \log_2 n$, dividing by n and letting $n \rightarrow \infty$ gives the claimed result. \blacksquare

Note that we have actually proved a stronger result, namely,

$$\mathbb{E} \left[\left(\ell(f_n^*(X^n)) - \imath_{X^n}(X^n) \right)^2 \right] = O(\log^2 n). \quad (221)$$

Linear information growth is sufficient for dispersion to equal varentropy:

Theorem 21: If the source has linear information growth, and finite varentropy, then:

$$D = \sigma^2. \quad (222)$$

Proof: For notational convenience, we abbreviate H_n for $H(X^n)$. Expanding the definition of the variance of ℓ_n , we obtain,

$$\begin{aligned} \text{Var}(\ell(f_n^*(X^n))) &= \mathbb{E}[(\ell_n - \mathbb{E}[\ell_n])^2] \\ &= \mathbb{E}[(\ell_n - \imath_n) + (\imath_n - H_n) - \mathbb{E}[\ell_n - \imath_n]]^2 \end{aligned} \quad (223)$$

$$= \mathbb{E}[(\ell_n - \imath_n) + (\imath_n - H_n) - \mathbb{E}[\ell_n - \imath_n]]^2 \quad (224)$$

$$= \mathbb{E}[(\ell_n - \imath_n)^2] + \mathbb{E}[(\imath_n - H_n)^2] - \mathbb{E}^2[\ell_n - \imath_n] + 2\mathbb{E}[(\ell_n - \imath_n)(\imath_n - H_n)] \quad (225)$$

and therefore, using the Cauchy-Schwarz inequality twice,

$$\begin{aligned} & |\text{Var}(\ell(f_n^*(X^n))) - \text{Var}(\iota_{X^n}(X^n))| \\ &= |\mathbb{E}[(\ell_n - \iota_n)^2] - \mathbb{E}^2[\ell_n - \iota_n] + 2\mathbb{E}[(\ell_n - \iota_n)(\iota_n - H_n)]| \end{aligned} \quad (226)$$

$$\leq 2\mathbb{E}[(\ell_n - \iota_n)^2] + 2\left\{\mathbb{E}[(\ell_n - \iota_n)^2]\right\}^{1/2} [\text{Var}(\iota_{X^n}(X^n))]^{1/2}. \quad (227)$$

Dividing by n and letting $n \rightarrow \infty$, we obtain that the first term tends to zero by Lemma 1, and the second term becomes,

$$2\left\{\frac{\mathbb{E}[(\ell_n - \iota_n)^2]}{n}\right\}^{1/2} \left\{\frac{\text{Var}(\iota_{X^n}(X^n))}{n}\right\}^{1/2}, \quad (228)$$

which also tends to zero by Lemma 1 and the finite-varentropy rate assumption. Therefore,

$$\lim_{n \rightarrow \infty} \frac{1}{n} |\text{Var}(\ell(f_n^*(X^n))) - \text{Var}(\iota_{X^n}(X^n))| = 0, \quad (229)$$

which, in particular, implies that $\sigma^2 = D$. \blacksquare

In view of (221), if we normalize by $\sqrt{n} \log n$, instead of n in the last step of the proof of Theorem 21, we obtain the stronger result:

$$|\text{Var}(\ell(f_n^*(X^n))) - \text{Var}(\iota_{X^n}(X^n))| = O(\sqrt{n} \log_2 n). \quad (230)$$

Also, Lemma 1 and Theorem 21 remain valid if instead of the linear information growth condition we invoke the weaker assumption that there exists a sequence $\tau_n = o(\sqrt{n})$, such that,

$$\max_{x^n: P_{X^n}(x^n) \neq 0} \iota_{X^n}(x^n) = o(2^{\epsilon_n/2}). \quad (231)$$

We turn now attention to the Markov chain case.

Theorem 22: Let $\{X_n\}$ be an irreducible, aperiodic (not necessarily stationary) Markov source with entropy rate H . Then:

- 1) The varentropy rate σ^2 defined in (108) exists as the limit,

$$\sigma^2 = \lim_{n \rightarrow \infty} \frac{1}{n} \text{Var}(\iota_{X^n}(X^n)). \quad (232)$$

- 2) The dispersion D defined in (208) exists as the limit,

$$D = \lim_{n \rightarrow \infty} \frac{1}{n} \text{Var}(\ell(f_n^*(X^n))). \quad (233)$$

- 3) $D = \sigma^2$.

- 4) The varentropy rate (or, equivalently, the dispersion) can be characterized in terms of the best achievable rate $R^*(n, \epsilon)$ as,

$$\sigma^2 = \lim_{\epsilon \rightarrow 0} \lim_{n \rightarrow \infty} \frac{n(R^*(n, \epsilon) - H)^2}{2 \ln \frac{1}{\epsilon}} = \lim_{\epsilon \rightarrow 0} \lim_{n \rightarrow \infty} n \left(\frac{R^*(n, \epsilon) - H}{Q^{-1}(\epsilon)} \right)^2, \quad (234)$$

as long as σ^2 is nonzero.

Proof: The limiting expression in part 1) was already established in Theorem 12 of Section IV; see also the discussion leading to (167) in Section VI. Recalling that every irreducible and aperiodic Markov source is a linear information growth source, combining part 1) with Theorem 21 immediately yields the results of parts 2) and 3).

Finally, part 4) follows from the results of Section VI. Under the present assumptions, Theorems 18 and 19 together imply that there is a finite constant C_1 such that,

$$\left| \sqrt{n}(R^*(n, \epsilon) - H) - \sigma Q^{-1}(\epsilon) \right| \leq \frac{1}{2} \frac{\log_2 n}{\sqrt{n}} + \frac{C_1}{\sqrt{n}}, \quad (235)$$

for all $\epsilon \in (0, 1/2)$ and all n large enough. Therefore,

$$\lim_{n \rightarrow \infty} n(R^*(n, \epsilon) - H)^2 = \sigma^2(Q^{-1}(\epsilon))^2. \quad (236)$$

Dividing by $2 \ln \frac{1}{\epsilon}$, letting $\epsilon \downarrow 0$, and recalling the simple fact that $(Q^{-1}(\epsilon))^2 \sim 2 \ln \frac{1}{\epsilon}$ (see, e.g., [28, Section 3.3]) proves (234) and completes the proof of the theorem. ■

From Theorem 21 it follows that, for a broad class of sources including all ergodic Markov chains with nonzero varentropy rate,

$$\lim_{n \rightarrow \infty} \frac{\text{Var}(\ell(f_n^*(X^n)))}{\text{Var}(\iota_{X^n}(X^n))} = 1. \quad (237)$$

Analogously to Theorem 9, we could explore whether (237) might hold under broader conditions, including the general setting of possibly non-serial sources. However, consider the following simple example.

Example 6: As in Example 3, let X_M be equiprobable on a set of M elements, then,

$$H(X_M) = \log_2 M \quad (238)$$

$$\text{Var}(\iota_{X_M}(X_M)) = 0 \quad (239)$$

$$\limsup_{M \rightarrow \infty} \text{Var}(\ell(f^*(X_M))) = 2 + \frac{1}{4} \quad (240)$$

$$\liminf_{M \rightarrow \infty} \text{Var}(\ell(f^*(X_M))) = 2. \quad (241)$$

To verify (240) and (241), define the function,

$$s(K) = \sum_{i=1}^K i^2 2^i \quad (242)$$

$$= -6 + 2^{K+1}(3 - 2K + K^2). \quad (243)$$

It is straightforward to check that,

$$\mathbb{E}[\ell^2(f^*(X_M))] = \frac{1}{M} (s(\lfloor \log_2 M \rfloor) - (\lfloor \log_2 M \rfloor)^2 \cdot (2^{\lfloor \log_2 M \rfloor + 1} - M - 1)). \quad (244)$$

Together with (90), (244) results in,

$$\text{Var}(\ell(f^*(X_M))) = 3\xi_M - \xi_M^2 + o(1), \quad (245)$$

with,

$$\xi_M = \frac{2^{1+\lceil \log_2 M \rceil}}{M}, \quad (246)$$

which takes values in $(1, 2]$. On that interval, the parabola $3x - x^2$ takes a minimum value of 2 and a maximum value of $(3/2)^2$, and (240), (241) follow.

Although the ratio of optimal codelength variance to the varentropy rate may be infinity as illustrated in Example 6, we do have the following counterpart of the first-moment result in Theorem 9 for the second moments:

Theorem 23: For any (not necessarily serial) source $\mathbf{X} = \{P_{X^{(n)}}\}_{n=1}^\infty$,

$$\lim_{n \rightarrow \infty} \frac{\mathbb{E}[\ell^2(\mathbf{f}_n^*(X^{(n)}))]}{\mathbb{E}[\imath_{X^{(n)}}^2(X^{(n)})]} = 1, \quad (247)$$

as long as the denominator diverges.

Proof: Theorem 2 implies that,

$$\mathbb{E}[\ell^2(\mathbf{f}_n^*(X^{(n)}))] \leq \mathbb{E}[\imath_{X^{(n)}}^2(X^{(n)})]. \quad (248)$$

Therefore, the lim sup in (247) is bounded above by 1. To establish the corresponding lower bound, fix an arbitrary $\vartheta > 0$. Then,

$$\mathbb{E}[\ell^2(\mathbf{f}_n^*(X^{(n)}))] = \sum_{k \geq 1} \mathbb{P}[\ell^2(\mathbf{f}_n^*(X^{(n)})) \geq k] \quad (249)$$

$$= \sum_{k \geq 1} \mathbb{P}[\ell_n^* \geq \sqrt{k}] \quad (250)$$

$$= \sum_{k \geq 1} \mathbb{P}[\ell_n^* \geq \lceil \sqrt{k} \rceil] \quad (251)$$

$$\geq \sum_{k \geq 1} \left[\mathbb{P}[\imath_{X^{(n)}}(X^{(n)}) \geq (1 + \vartheta) \lceil \sqrt{k} \rceil] - 2^{-\vartheta \lceil \sqrt{k} \rceil} \right], \quad (252)$$

where (252) follows by letting $\tau = \vartheta \lceil \sqrt{k} \rceil$ in the converse Theorem 3. Therefore,

$$\mathbb{E}[\ell^2(\mathbf{f}_n^*(X^{(n)}))] \geq -C_\vartheta + \sum_{k \geq 1} \mathbb{P}[\imath_{X^{(n)}}^2(X^{(n)}) \geq (1 + \vartheta)^2 \lceil \sqrt{k} \rceil^2] \quad (253)$$

$$\geq -D_\vartheta + \sum_{k \geq 1} \mathbb{P}\left[\frac{\imath_{X^{(n)}}^2(X^{(n)})}{(1 + \vartheta)^3} \geq k\right] \quad (254)$$

$$\geq -D_\vartheta + \frac{1}{(1 + \vartheta)^3} \mathbb{E}[\imath_{X^{(n)}}^2(X^{(n)})]. \quad (255)$$

where C_ϑ, D_ϑ are positive scalars that only vary with ϑ . Note that (253) holds because $a^{-\sqrt{k}}$ is summable for all $0 < a < 1$; (254) holds because $(1 + \vartheta)k \geq \lceil \sqrt{k} \rceil^2$ for all sufficiently large k ; and (255) holds because,

$$\int_k^{k+1} (1 - F(x)) dx \geq 1 - F(k + 1), \quad (256)$$

whenever $F(x)$ is a cumulative distribution function. Dividing both sides of (253)-(255) by the second moment $\mathbb{E}[t_{X^{(n)}}^2(X^{(n)})]$ and letting $n \rightarrow \infty$, we conclude that the ratio in (247) is lower bounded by $(1 + \vartheta)^{-3}$. Since ϑ can be taken to be arbitrarily small, this proves that the \liminf (247) is lower bounded by 1, as required. ■

ACKNOWLEDGMENTS

The work of SV was supported in part by the National Science Foundation (NSF) under Grant CCF-1016625 and by the Center for Science of Information (CSoI), an NSF Science and Technology Center, under Grant CCF-0939370. The work of IK was supported in part by the research program ‘CROWN’ through the Operational Program ‘Education and Lifelong Learning 2007-2013’ of NSRF. Parts of this paper were presented at the 46th Annual Conference on Information Sciences and Systems, Princeton University, Princeton, NJ, March 21-23, 2012.

REFERENCES

- [1] N. Alon and A. Orlitsky, “A lower bound on the expected length of one-to-one codes,” *IEEE Trans. Information Theory*, vol. 40, pp. 1670-1672, Sep. 1994
- [2] A. R. Barron, “*Logically smooth density estimation*,” Ph. D. thesis, Dept. Electrical Engineering, Stanford University, Sep. 1985
- [3] B.C. Bradley, “Basic properties of strong mixing conditions,” in *Dependence in Probability and Statistics*, E. Wileln and M. S. Taqqu, Eds. Birkhäuser, Boston, 1986, pp. 165–192.
- [4] T. Cover and J. Thomas, *Elements of Information Theory*. 2nd Ed., New York: Wiley, 2006
- [5] K.L. Chung, *Markov Chains with Stationary Transition Probabilities*, Springer-Verlag, New York, 1967.
- [6] A. Dembo and I. Kontoyiannis, “Source coding, large deviations, and approximate pattern matching,” *IEEE Trans. Inform. Theory*, vol. 48, pp. 1590–1615, June 2002.
- [7] P. Hall, *Rates of Convergence in the Central Limit Theorem*, Pitman Advanced Publishing Program, London, 1982
- [8] I. A. Ibragimov, “Some limit theorems for stationary processes,” *Theory Probab. Appl.*, vol. 7, pp. 349–382, 1962.
- [9] M. Hayashi, “Second-Order Asymptotics in Fixed-Length Source Coding and Intrinsic Randomness,” *IEEE Trans. Inform. Theory*, vol. 54, no. 10, pp. 4619-4637, October 2008.
- [10] J.C. Kieffer, “Sample converses in source coding theory,” *IEEE Trans. on Inform. Theory*, vol. IT-37, no. 2, pp. 263–268, 1991.
- [11] I. Kontoyiannis, “Second-order noiseless source coding theorems,” *IEEE Trans. Inform. Theory*, vol. 43, no. 3, pp. 1339-1341, July 1997.
- [12] I. Kontoyiannis, “Asymptotic recurrence and waiting times for stationary processes,” *J. Theoret. Probab.*, vol. 11, pp. 795-811, 1998.
- [13] I. Kontoyiannis and S.P. Meyn, “Spectral theory and limit theorems for geometrically ergodic Markov processes,” *Ann. Appl. Probab.*, vol. 13, pp. 304-362, 2003.
- [14] I. Kontoyiannis and Yu.M. Suhov, “Prefixes and the entropy rate for long-range sources.” Chapter in *Probability Statistics and Optimization*, (F.P. Kelly, ed.). Wiley, New York, 1994.
- [15] V. Yu. Korolev and I. G. Shevtsova, “On the upper bound for the absolute constant in the Berry-Esséen inequality,” *Theory of Probability and its Applications*, vol. 54, no. 4, pp. 638-658.
- [16] B. Mann, *Berry-Esséen Central Limit Theorems for Markov Chains*, PhD thesis, Department of Mathematics, Harvard University, 1996.
- [17] B. McMillan, “The basic theorems of information theory,” *Ann. Math. Statist.*, vol. 24, pp. 196219, June 1953.
- [18] B. McMillan, “Two inequalities implied by unique decipherability,” *IRE Trans. Inform. Theory*, vol. IT-2, pp. 115116, Dec. 1956.
- [19] S.P. Meyn and R.L. Tweedie, *Markov Chains and Stochastic Stability*, second edition, Cambridge University Press, 2009.
- [20] S.V. Nagaev, “More exact limit theorems for homogeneous Markov chains,” *Theory Probab. Appl.*, vol. 6, pp. 62-81, 1961.
- [21] V. V. Petrov, *Limit Theorems of Probability Theory: Sequences of Independent Random Variables*, Oxford Science Publications, 1995

- [22] W. Philipp and W. Stout, *Almost Sure Invariance Principles for Partial Sums of Weakly Dependent Random Variables*, Memoirs of the AMS, 1975.
- [23] Y. Polyanskiy, H. V. Poor and S. Verdú, “Channel coding rate in the finite blocklength regime,” *IEEE Trans. Inform. Theory*, vol. 56, no. 5, pp. 2307–2359, May 2010
- [24] C. E. Shannon, “A mathematical theory of communication,” *Bell Syst. Tech. J.*, vol. 27, pp. 379–423 and 623–656, July and October, 1948.
- [25] P.C. Shields, *The Ergodic Theory of Discrete Sample Paths*, Graduate Studies in Mathematics, American Mathematical Society, 1996.
- [26] V. Strassen, “Asymptotische Abschätzungen in Shannons Informationstheorie,” *Trans. Third Prague Conf. Information Theory, on Statistics, Decision Functions, Random Processes* (Liblice, 1962), pages 689–723., Publ. House Czech. Acad. Sci., Prague, 1964.
- [27] W. Szpankowski and S. Verdú, “Minimum Expected Length of Fixed-to-Variable Lossless Compression without Prefix Constraints,” *IEEE Trans. on Information Theory*, vol. 57, no. 7, pp. 4017–4025, July 2011.
- [28] S. Verdú, *Multiuser Detection*, Cambridge University Press, 1998.
- [29] S. Verdú, *EE528—Information Theory, Lecture Notes*, Princeton University, Princeton, NJ, 2011.
- [30] S. Verdú, “teaching it,” XXVIII Shannon Lecture, *2007 IEEE International Symposium on Information Theory*, Nice, France, June 28, 2007.
- [31] S. Verdú, “Teaching Lossless Data Compression,” *IEEE Information Theory Society Newsletter*, vol. 61, no. 1, pp. 18–19, April 2011
- [32] A. D. Wyner, “An Upper Bound on the Entropy Series,” *Inform. Control*, 20, 176–181, 1972.
- [33] A. A. Yushkevich, “On limit theorems connected with the concept of entropy of Markov chains”, *Uspekhi Matematicheskikh Nauk*, 8:5(57), pp. 177–180, 1953